DOCUMENT RESUME

ED 235 198 TM 830 595

AUTHOR Cook, Linda L.; Eignor, Daniel R.
TITLE An Investigation of the Feasibility of Applying Item Response Theory to Equate Achievement Tests.
SPONS AGENCY Educational Testing Service, Princeton, N.J.
PUB DATE Apr 83
NOTE 75p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, Montreal, Quebec, April 11-15, 1983). This study was supported through Program Research Planning Council funding.
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Achievement Tests; *College Entrance Examinations; Comparative Analysis; *Equated Scores; Feasibility Studies; Goodness of Fit; *Latent Trait Theory
IDENTIFIERS College Board Achievement Tests; *Equipercentile Equating; Graduate Record Examinations; *Linear Equating Method

ABSTRACT

The purpose of this study was to examine the feasibility of using item response theory (IRT) methods to equate different forms of three College Board Achievement Tests (Biology, American History and Social Studies, and Mathematics Level II) and one Graduate Record Examinations Achievement Test (Advanced Biology), rather than conventional or equipercentile methods. The criterion for evaluation of the results was scale drift, which is said to have occurred if the results of equating test form A directly to test form D is not the same as that obtained by equating test form A to test form D through intervening forms B and C. The results of three conventional linear equating methods, conventional equipercentile equating with an anchor test, and two IRT equating methods were compared. No linear equating method produced scaled scores that could be considered seriously discrepant from the criterion scores, indicating that they perform quite adequately. The equipercentile method produced the largest total error. The IRT concurrent and characteristic curve transformation methods gave very similar results, and results indicate that it is feasible to use IRT to equate the tests in this study. (BW)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED235198

# An Investigation of the Feasibility of Applying Item Response Theory to Equate Achievement Tests[1,2,3,4]

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Linda L. Cook
Daniel R. Eignor

Educational Testing Service

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
L. L. Cook
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 830 595

[1]A paper presented at the annual meeting of AERA, Montreal, 1983.

[2]This study was supported by Educational Testing Service through Program Research Planning Council funding.

[3]The authors would like to acknowledge the assistance of Nancy Ervin and Karen Carroll in performing this study.

[4]The Appendix to this paper is available upon request from the first author, at the following address: Educational Testing Service, Mail Stop 31-E, Rosedale Rd. Princeton, NJ 08541.

# An Investigation of the Feasibility of Applying Item Response Theory to Equate Achievement Tests

Linda L. Cook
Daniel R. Eignor

Educational Testing Service

## Introduction

Most admissions testing programs develop and administer many different forms (versions) of the same test. They typically do so in order to ensure equity to examinees whose scores on the different forms of the test may be compared. The reason multiple forms of the same test are necessary, if equity is to be provided for all examinees who take the test, becomes apparent if one considers the following situation. Suppose that only a single form of the Scholastic Aptitude Test (SAT) was administered at each of the many test administrations that occur in a year. Examinees taking the test at the end of the year would most surely have some prior knowledge of the items on the test and would have a definite advantage over examinees who took the test at the beginning of the year.

Because equity is a major concern of admissions testing programs, different forms of a test administered by a program are constructed to be as similar in difficulty and content as possible so that a particular examinee will not be advantaged simply because he/she took an easier version of the test. Unfortunately, in spite of efforts on the part of those constructing the test forms, it is usually impossible to assemble different forms of the same test that are of exactly the same difficulty level. Therefore it becomes necessary, if the testing program is to

accomplish its goal of equity, to establish a process that renders scores on the different test forms comparable. This process, which is referred to as equating, provides a transformation of raw scores to scaled scores (scores on an arbitrarily chosen common scale). Ideally, the end result of the equating process is that an examinee would receive the same scaled score regardless of which form of the test he/she was administered.

There exist many different data collection designs and equating models that can be used to establish a transformation of raw scores on a particular form of a test to scaled scores. Whether or not the equating process is effective (the testing program's goal of equity is realized) depends largely on how well the data collected fit the underlying assumptions of the particular equating model as well as how robust the specific model is to violations of these assumptions.

The purpose of this study was to examine the feasibility of using item response theory (IRT) methods to equate different forms of three achievement tests (Biology, American History and Social Studies, and Mathematics Level II) that are administered by the College Board Admissions Testing Program and one achievement test (Advanced Biology) that is part to the Graduate Record Examinations Achievement Test battery. All of the tests investigated in this study are typically equated using conventional linear or curvilinear (equipercentile) methods. It was considered important to investigate the possibility of using IRT to equate these tests because, if the type of data collected from administrations of the tests fit an IRT model, or if the particular IRT equating model is sufficiently robust, a number of advantages accrue; these include:

1. An improved method for curvilinear equating. When test forms that differ considerably in level of difficulty are equated to each other, the relationship between raw scores on the two forms is typically curvilinear. Conventional linear equating methods cannot reflect this curvilinear relationship. On the other hand, conventional equipercentile methods, while reflecting the curvilinearity of the relationship, often lead to unstable equating of extreme scores because of scarcity of data in the tails of the score distribution.
2. Easier re-equating should it be decided not to score an item after a particular form of the test has been administered. Conventional equating methods require that the shortened test be rescored. This is not necessary when using IRT equating methods.
3. The possible reduction in scale drift which may occur when less robust equating methods are used over time, most notably when the test forms are not parallel and the equating samples differ in level of ability.
4. The possibility of pre-equating, or deriving the relationship between scores on the two test forms before they are administered. This is possible only when items have been pretested. The use of IRT for pre-equating offers a unique contribution that is impossible to obtain using conventional equating methods.

As mentioned previously, in order for the above listed advantages of IRT equating to accrue, the data collected for the equating must meet the

underlying assumptions of the particular IRT model or the model must be sufficiently robust to violations of these assumptions when used for equating purposes. A fundamental assumption underlying all IRT models, is unidimensionality, i.e. a test should measure only a single trait or ability. Whether or not this assumption is met by achievement test data is questionable. It is quite likely that tests that have been constructed to measure a variety of specific content areas (typically the case for achievement tests) will yield data of a multidimensional nature. One way to investigate the feasibility of using IRT methods to equate achievement tests is to compare the results of IRT equating to results obtained from conventional methods that have gained credibility through a long period of use for equating these tests.

## Overview of the Study

A problem related to evaluation of the results of any equating method concerns the choice of a criterion measure. Since it is impossible to determine what the true equating should be, i.e. the true criterion against which to judge the actual equating, other criterion measures have often been devised, these vary in degree of complexity and assumptions made (see Cook and Eignor, 1983, for a review of some of the more commonly used criteria for equating studies). The criterion used in the present study was scale drift. This criterion was used successfully in a study by Petersen, Cook and Stocking (in press), which compared the results of using IRT and conventional equating methods to equate the verbal and mathematical

sections of the SAT. Scale drift is said to have occurred if the results of equating test form A directly to test form D is not the same as that obtained by equating test form A to test form D through intervening forms B and C. In order to evaluate scale drift for the four achievement tests investigated in this study, a closed circular chain of equatings was formed for each of the tests. Figure 1 contains a diagram of the four equating chains. Upper case letter and number combinations indicate particular achievement test forms and the abbreviation CI indicates common items linking adjacent achievement test forms. It is possible to use the equating chains shown in Figure 1 to equate a test form to itself through a number of intervening test forms. If no scale drift has occurred, the initial (criterion) and final scaled scores for the forms should be identical. Any discrepancy between initial and final scores for a test form is attributed to scale drift resulting from application of the particular equating method.

Scale drift was used as the criterion to compare the results of three conventional linear equating methods (Tucker, Levine Equally Reliable and Levine Unequally Reliable), conventional equipercentile equating with an anchor test, and two IRT equating methods. The two IRT methods are referred to as (1) the concurrent method and (2) the characteristic curve transformation method. The results of the various equating methods were compared both graphically and analytically.

In addition to the evaluation of the equating results, an effort was made to assess the goodness of fit of the individual achievement test items

ATP American History and Social Studies Test

AAC → CI → XAC → CI → K-UAC2 → CI
↑ ↓
CI ← YAC1 ← CI ← K-WAC ← CI ← YAC2

ATP Math Level II Test

CAC2 → CI → WAC → CI → AAC → CI → VAC1
↑ ↓
CI ← BAC ← CI ← ZAC ← CI ← XAC ← CI

ATP Biology Test

BAC → CI → UAC2 → CI → XAC → CI → TAC2 → CI
↑
YAC ← CI ← WAC ← CI ← UAC1 ← CI ← SAC2 ← CI ← VAC1

GRE Advanced Biology Test

SGR → CI → K2-UGR1 → CI → WGR → CI
↑ ↓
CI ← K-UGR2 ← CI ← XGR ← CI ← ZGR

Figure 1: ATP and GRE Achievement Test equating chains. Letters and letter-number combinations indicate achievement test forms. The abbreviation CI is used to indicate common items shared by two test forms.

to the IRT model used for this study. The goodness of fit assessment was carried out using a chi-square like statistic, referred to as $Q_1'$, in conjunction with item ability regression plots. The statistic and the plots are described in the methodology section of this paper.

## Methodology

### Description of the Tests

As mentioned previously, three of the achievement tests used in this study (Biology, Mathematics Level II, and American History and Social Studies) are administered by the College Board Admissions Testing Program (ATP). The fourth achievement test is the Advanced Biology Test administered by the Graduate Record Examinations (GRE) program. The Admissions Testing Program Achievement Tests are multiple choice tests that are used in conjunction with measures of high school performance, as well as other standardized tests such as the SAT, to select students for admission to colleges and universities. The Graduate Record Examinations Subject (Advanced) Tests are also multiple choice tests that are designed to help graduate school committees and fellowship sponsors assess the qualifications of applicants for advanced study and for fellowship awards. The GRE Program recommends that scores on Advanced Tests be used in conjunction with other relevant information when making admissions or award decisions.

The ATP Biology and American History and Social Studies Tests are 60 minute tests that each contain 100 items. The ATP Biology Test covers the

following topics: cellular structure and function; organismal reproduction, development, growth, nutrition, structure, and function; genetics; evolution; systematics; ecology; and behavior. The test also includes questions that require the interpretation of experimental data, understanding of scientific methods and laboratory techniques, and knowledge of the history of biology. Questions on the ATP American History and Social Studies Test emphasize history from the nineteenth and twentieth centuries rather than earlier time periods. The fields of American History that are examined are political, social, economic, diplomatic, intellectual, and cultural history. Political history receives the most attention, social and economic history somewhat less; intellectual and cultural history receives the least attention. The ATP Mathematics Level II test contains 50 items and is also administered in a 60 minute time period. The test is composed of approximately equal parts of algebra, geometry, trigonometry, functions, and a miscellaneous category consisting of such topics as sequences and limits, logic and proof, probability and statistics, and number theory.

Specifications for the GRE Advanced Biology Test have changed somewhat over the past few years. Of the test forms that comprise the GRE Biology equating chain used in this study, Form SGR contains 200[1] items and was administered with a 180 minute time limit. The other test forms each contain 210[1] items and were administered with a 170 minute time limit. The

---

[1]GRE Biology Forms SGR and XGR each contain one item that was not scored for score reporting purposes.

items in all of the forms comprising the GRE Biology equating chain are assigned to three non-overlapping subscores. The subscores for Form SGR were used for experimental purposes only. Subscores for the remaining forms in the chain are actually used for score reporting. The subscores are referred to as: (1) Cellular and Subcellular Biology; (2) Organismal Biology; and (3) Population Biology. Each subscore covers a fairly wide range of content that can be classified under these general headings.

Raw scores on the ATP Achievement Tests are typically transformed to scaled scores on a 200 to 800 scale via linear equating methods. Linear equating methods are also typically used to transform GRE Achievement Test raw scores to a 200 to 990 scale. Raw scores on all the tests used in this study are obtained scores that have been corrected for guessing. Raw scores are computed by the formula $R-(1/k)(W)$, where R is the number of correct responses, W is the number of incorrect responses, and $(k+1)$ is the number of choices per item.

Data Collection

All equating methods have two components, a design for data collection and a statistical model for analyzing the data. An internal anchor test design (Angoff, 1971) was used in this study for data collection. An anchor test design requires administering one form of a test to one group of examinees, a second form to a second group of examinees, and a common set of items (anchor test) to both groups. The anchor test may be included within the total test (internal anchor) or it may be administered separately (external anchor).

Two samples (which varied in size from approximately 2,000 to approximately 4,000 cases) were randomly selected for each test form used in this study. Whenever possible, samples for the experimental equatings were selected from the same populations (test administrations) used when the test forms were originally introduced and placed on scale. Table 1 contains descriptive information regarding the samples. The table includes raw-score summary statistics for the total test and anchor test as well as dates of the test administrations from which the samples were selected.

Criterion

In order to assess the magnitude of scale drift associated with an equating method, each test form in a chain (see Figure 1) was equated to the preceding form. For example, for the ATP American History and Social Studies chain, Form AAC was treated as the initial form of the test in the chain. For each equating method used in the study, the raw to scale transformation obtained from equating Form AAC to itself through the five intervening forms was compared to the initial AAC scale. Any discrepancy between the raw to scale transformation obtained from the circular chain of equatings and the initial AAC scale was considered to be scale drift attributable to the equating method.

Conventional Equating Methods

The conventional curvilinear equating method used in this study was equipercentile equating. Equipercentile equating is based on the principle that scores on two test forms given to the same group of examinees will be

Table 1

Raw Score[a] Summary Statistics for Equating Samples

| Form | Admin. Date | N | Total Test | | | Anchor Test | | | Anchor Test/Total Test Correlation |
|---|---|---|---|---|---|---|---|---|---|
| | | | n | $\bar{X}$ | SD | n | $\bar{X}$ | SD | |
| **ATP Biology Test** | | | | | | | | | |
| 3BAC | 12/79 | 2309 | 100 | 49.87 | 18.69 | 20 | 10.68 | 4.67 | .87 |
| UAC2 | 1/78 | 2699 | 100 | 54.47 | 19.27 | 20 | 12.19 | 4.68 | .88 |
| UAC2 | 1/76 | 2394 | 100 | 46.67 | 19.54 | 20 | 9.21 | 4.85 | .87 |
| XAC | 1/75 | 2042 | 100 | 46.88 | 19.84 | 20 | 8.97 | 4.78 | .87 |
| XAC | 3/77 | 2314 | 100 | 45.11 | 19.77 | 20 | 9.43 | 4.86 | .87 |
| TAC2 | 1/78 | 2511 | 100 | 43.75 | 18.70 | 20 | 9.59 | 4.77 | .86 |
| TAC2 | 5/79 | 3032 | 100 | 47.59 | 19.88 | 20 | 10.64 | 4.56 | .89 |
| VAC1 | 1/73 | 2101 | 100 | 43.70 | 17.94 | 20 | 10.00 | 4.39 | .87 |
| VAC1 | 5/78 | 3253 | 100 | 48.38 | 18.77 | 20 | 9.88 | 4.40 | .86 |
| SAC2 | 11/77 | 3344 | 100 | 48.89 | 19.60 | 20 | 10.18 | 4.29 | .85 |
| SAC2 | 11/77 | 3344 | 100 | 48.89 | 19.60 | 20 | 10.98 | 4.64 | .88 |
| UAC1 | 11/76 | 3732 | 100 | 51.86 | 20.00 | 20 | 10.91 | 4.68 | .90 |
| UAC1 | 1/79 | 2259 | 100 | 47.06 | 19.05 | 27 | 14.02 | 6.03 | .90 |
| WAC | 1/74 | 2019 | 100 | 45.13 | 19.51 | 27 | 12.95 | 6.14 | .91 |
| WAC | 12/75 | 2064 | 100 | 48.01 | 19.81 | 26 | 13.19 | 5.93 | .91 |
| YAC | 12/76 | 2129 | 100 | 51.89 | 18.25 | 26 | 13.05 | 5.53 | .90 |
| YAC | 12/76 | 2129 | 100 | 51.89 | 18.25 | 20 | 9.32 | 4.31 | .85 |
| 3BAC | 12/79 | 2309 | 100 | 49.87 | 18.69 | 20 | 9.67 | 4.37 | .85 |
| **GRE Biology Test** | | | | | | | | | |
| SGR | 12/70 | 3214 | 199 | 88.97 | 25.78 | 72 | 34.37 | 11.18 | .94 |
| K2-UGR1 | 4/75 | 2086 | 210 | 92.69 | 27.85 | 72 | 33.31 | 10.74 | .92 |
| K2-UGR1 | 6/76 | 2039 | 210 | 94.60 | 29.74 | 69 | 33.47 | 11.39 | .94 |
| WGR | 10/74 | 2153 | 210 | 97.05 | 28.96 | 69 | 34.57 | 10.84 | .93 |
| WGR | 10/74 | 2153 | 210 | 97.05 | 28.96 | 47 | 21.08 | 7.69 | .89 |
| ZGR | 12/77 | 2294 | 210 | 103.58 | 30.35 | 47 | 21.05 | 7.93 | .89 |
| ZGR | 10/78 | 1966 | 210 | 104.19 | 29.68 | 45 | 23.12 | 7.20 | .88 |
| XGR | 1/78 | 3320 | 209 | 101.07 | 27.06 | 45 | 22.73 | 6.85 | .89 |
| XGR | 12/75 | 2351 | 209 | 101.81 | 28.76 | 68 | 38.89 | 11.06 | .94 |
| K-UGR2 | 10/74 | 2012 | 210 | 92.01 | 30.25 | 68 | 38.54 | 10.97 | .93 |
| K-UGR2 | 10/74 | 2012 | 210 | 92.01 | 30.25 | 55 | 28.59 | 9.22 | .92 |
| SGR | 12/70 | 3214 | 199 | 88.97 | 25.78 | 55 | 29.08 | 9.32 | .93 |

[a]Raw scores are obtained scores that have been corrected for guessing.

Table 1 (continued)

Raw Score[a] Summary Statistics for Equating Samples

| Form | Admin. Date | N | Total Test | | | Anchor Test | | | Anchor Test/Total Test Correlation |
|---|---|---|---|---|---|---|---|---|---|
| | | | n | $\bar{X}$ | SD | n | $\bar{X}$ | SD | |
| **ATP Mathematics Level II Test** | | | | | | | | | |
| 3CAC2 | 12/80 | 2117 | 50 | 24.49 | 9.63 | 17 | 8.59 | 3.73 | .90 |
| WAC | 1/74 | 2160 | 50 | 22.84 | 10.71 | 17 | 7.86 | 4.07 | .92 |
| WAC | 4/76 | 1917 | 50 | 21.47 | 11.14 | 15 | 7.27 | 4.17 | .92 |
| 3AAC | 12/78 | 2209 | 50 | 25.15 | 10.09 | 15 | 8.37 | 3.74 | .91 |
| 3AAC | 1/80 | 2343 | 50 | 24.56 | 10.42 | 15 | 7.69 | 3.59 | .91 |
| VAC1 | 1/73 | 2406 | 50 | 23.61 | 11.09 | 15 | 7.72 | 3.72 | .92 |
| VAC1 | 1/73 | 2406 | 50 | 23.61 | 11.09 | 19 | 9.96 | 4.59 | .93 |
| XAC | 1/75 | 2045 | 50 | 23.75 | 10.57 | 19 | 10.03 | 4.67 | .93 |
| XAC | 1/76 | 2025 | 50 | 24.04 | 10.60 | 20 | 9.70 | 4.29 | .93 |
| ZAC | 12/77 | 2081 | 50 | 23.82 | 9.64 | 20 | 9.91 | 3.88 | .91 |
| ZAC | 1/79 | 2600 | 50 | 22.92 | 10.27 | 20 | 9.22 | 4.57 | .93 |
| 3BAC | 12/79 | 2278 | 50 | 25.35 | 9.23 | 20 | 9.83 | 4.23 | .92 |
| 3BAC | 12/79 | 2278 | 50 | 25.35 | 9.23 | 17 | 8.73 | 3.40 | .90 |
| 3CAC2 | 12/80 | 2117 | 50 | 24.49 | 9.63 | 17 | 8.63 | 3.58 | .90 |
| **ATP American History and Social Studies Test** | | | | | | | | | |
| 3AAC | 12/78 | 2102 | 100 | 40.30 | 16.60 | 20 | 9.06 | 4.16 | .85 |
| XAC | 1/75 | 2058 | 100 | 33.97 | 15.54 | 20 | 8.72 | 4.32 | .86 |
| XAC | 4/76 | 2182 | 100 | 33.45 | 15.48 | 20 | 6.89 | 3.73 | .85 |
| K-UAC2 | 5/77 | 2554 | 100 | 37.69 | 17.67 | 20 | 7.31 | 3.93 | .85 |
| K-UAC2 | 5/77 | 2554 | 100 | 37.69 | 17.67 | 20 | 7.92 | 4.47 | .88 |
| YAC2 | 12/76 | 2120 | 100 | 38.73 | 15.13 | 20 | 7.28 | 4.14 | .84 |
| YAC2 | 1/79 | 2317 | 100 | 37.18 | 15.18 | 20 | 6.35 | 3.88 | .83 |
| K-WAC | 12/75 | 2144 | 100 | 30.16 | 17.03 | 20 | 6.81 | 4.12 | .86 |
| K-WAC | 5/79 | 2005 | 100 | 30.96 | 17.48 | 20 | 6.98 | 4.51 | .87 |
| YAC1 | 3/77 | 2141 | 100 | 37.87 | 16.48 | 20 | 6.53 | 4.42 | .86 |
| YAC1 | 6/76 | 2055 | 100 | 46.00 | 18.01 | 20 | 8.00 | 4.55 | .89 |
| 3AAC | 6/80 | 2031 | 100 | 46.93 | 17.92 | 20 | 9.08 | 4.42 | .87 |

[a]Raw scores are obtained scores that have been corrected for guessing.

considered equivalent if they correspond to the same percentile rank; i.e. equipercentile equating attempts to bring into coincidence the raw score distributions on two forms of a test given to the same group of examinees. Equipercentile equating is generally accomplished by setting equal scores on two test forms that have the same percentile rank in some group of examinees. Several different methods of equipercentile equating exist for application to anchor test equating designs (Angoff, 1971). The method used in this study actually requires two separate equipercentile equatings for each pair of forms in a particular equating chain. For example, in order to equate ATP Mathematics Level II scores on Form WAC to scores on Form CAC2, scores on CAC2 were set equal to scores on the common anchor test that have the same percentile rank for the group of examinees who took Form CAC2. The procedure was repeated for Form WAC, using the frequency distribution of scores for examinees who took Form WAC. Finally, scores on Forms CAC2 and WAC that correspond to the same score on the common anchor test (after the individual equipercentile equatings were accomplished) were said to be equivalent.

When applying equipercentile methods, some practitioners choose to smooth either the frequency distributions used in the equating or the resulting curve obtained from the equating. Because there is some controversy regarding when and how to smooth (e.g. see Angoff, 1971, p. 571), the authors chose to avoid confounding the equipercentile equating results with selection of a smoothing procedure. Thus, no smoothing was used at any point in the process.

The linear equating models used in this study were the Tucker, Levine Equally Reliable, and Levine Unequally Reliable models (Angoff, 1971). Linear equating methods assume that the score distributions on the two test forms to be equated differ only in their means and standard deviations. If this assumption is satisfied, a linear transformation will bring the score scales for the two forms into correspondence. However, if the distributions differ in more than their first and second moments, a more complex transformation (i.e. one provided by curvilinear equating methods such as equipercentile or IRT) will be needed to provide adequate equating of scores on the two test forms.

Linear equating methods all produce an equating transformation of the form $T(x) = Ax + B$, where T is the equating transformation, x is the test score to which it is applied, and A and B are parameters estimated from the data. The parameters A and B of the equating transformation are estimated by means of an equation that expresses the relationship between raw scores on two test forms in standard score terms:

$$(x-m_x)/s_x = (y-m_y)/s_y \ , \qquad (1)$$

where x and y refer to the test scores to be equated, and m and s refer to the means and standard deviations of the scores in some group of examinees. Methods using equation (1) differ in their identification of the means and standard deviations to be estimated. The Tucker and Levine Equally Reliable methods are based on the estimated means and standard deviations of observed scores whereas the Levine Unequally Reliable method is based on

the estimated means and standard deviations of true scores. For all three linear models, scores on the anchor test (common items) were used to estimate performance of the combined group of examinees on both the old and new forms of the test, thus simulating by statistical methods the situation in which the same group of examinees takes both forms of the test.

IRT Parameter Estimation

Item response theory assumes that there is a mathematical function which relates the probability of a correct response on an item to an examinee's ability. (See Lord, 1980, for a detailed discussion.) Many different mathematical models of this functional relationship are possible. The model chosen for this study was the three-parameter logistic model. In this model, where $\theta$ represents an examinee's ability, the probability of a correct response to item i, $P_i(\theta)$, is

$$P_i(\theta) = c_i + \frac{1-c_i}{1+e^{-1.702a_i(\theta-b_i)}} , \qquad (2)$$

where $a_i$, $b_i$, and $c_i$ are three parameters describing the item. These parameters have specific interpretations: $b_i$ is the point on the $\theta$ metric at the inflection point of $P_i(\theta)$ and is interpreted as the item difficulty; $a_i$ is proportional to the slope of $P_i(\theta)$ at the point of inflection and represents the item discrimination; and $c_i$ is the lower asymptote of $P_i(\theta)$ and represents a pseudo-guessing parameter.

The item parameters and examinee abilities for this study were estimated (calibrated) using the program LOGIST (Wingersky, Barton, and Lord, 1982; Wingersky, 1983). The estimates are obtained by a (modified) maximum likelihood procedure with special procedures for the treatment of omitted items (see Lord, 1974).

LOGIST produces as output estimates of the a, b, and c for each item, and $\theta$ for each examinee. The metric chosen arbitrarily for the $\theta$ (and b) scale is such that the distribution of estimates of $\theta$ has mean zero and standard deviation one. If two separate LOGIST runs are made for the same items, but different groups of examinees, the resulting parameter estimates will be on different scales. Theoretically, there is a simple linear relationship that transforms one scale to the other.

IRT Equating

One of the basic underlying properties of IRT that makes it useful for equating applications is the following. If the data being considered for the equating fit the assumptions of an IRT model, it is possible to obtain an estimate of an examinee's ability ($\theta$) that is independent of the items (test form) that the examinee responds to. Hence, it does not matter if an examinee takes an easy or hard form of a test; his/her ability estimate obtained from both forms will be identical, except for stochastic variation, once the parameter estimates obtained for the individual items are placed on the same scale. Further, if one is willing to use the ability ($\theta$) metric for score reporting purposes, IRT eliminates the need for equating different forms of a test; the only problem that requires

consideration is the placement of item parameter estimates, derived from independent calibrations, on the same scale.

For a variety of reasons, established testing programs (such as those whose data were used for this study) are often unable to report scores using the $\theta$ metric, and instead must continue to report scaled scores in a traditional manner, even though IRT has been used for equating purposes. Fortunately, because any ability score can be mathematically related to an estimated true score, it is possible to use ability scores to establish the relationship between (equate) estimated true scores on two forms of a test and subsequently transform the resulting equated true scores to traditional scaled scores.

The principle difference between the two IRT equating methods used in this study is derived from the manner in which the item parameter estimates were placed on the same scale prior to establishing the relationship between estimated true scores on two forms of a test. As mentioned previously, two IRT equating methods were studied; (1) the concurrent method and (2) the characteristic curve transformation method. For the concurrent method, each pair of achievement test forms (e.g. ATP Biology Forms BAC and UAC2) is calibrated in a single LOGIST run (see Figure 2). This results in item parameters on a common scale for each pair of test forms, represented by a box in Figure 2 (e.g., parameters for ATP Biology Forms BAC and UAC2 are on the same scale). However, each <u>separate</u> box shown in Figure 2 represents a unique scale (e.g., parameters for ATP Biology Form UAC2 which was calibrated with Form BAC are not on the same scale as parameters for Form UAC2 calibrated with Form XAC).

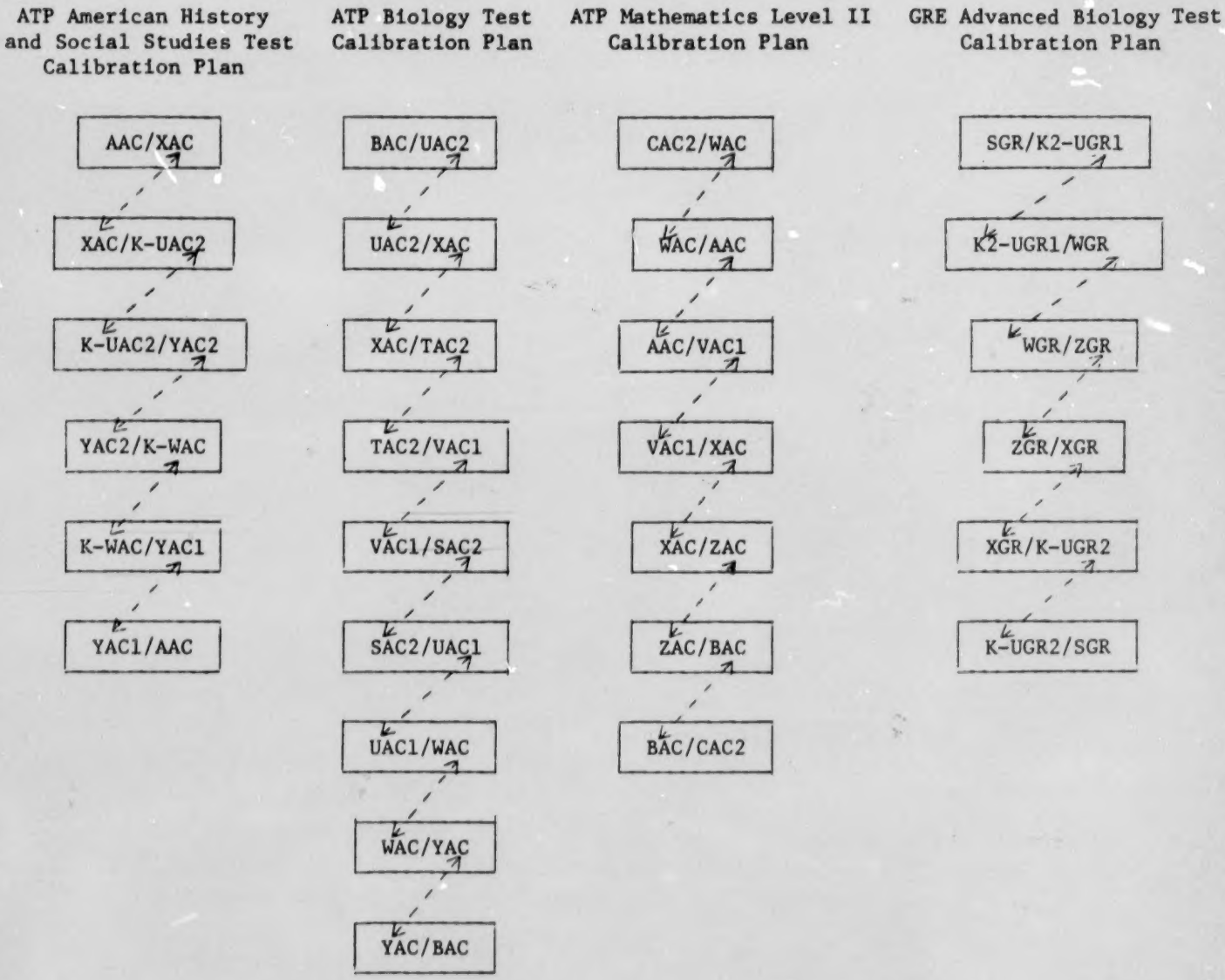



Figure 2: Achievement Test Calibration Plan. Boxes indicated separate calibration runs. Each box represents a sample of approximately 4000 examinees (2000 examinees who took the new form of the test and 2000 examinees who took the old form of the test. Dotted lines and arrows indicate common test forms that were used to place item parameter estimates from the separate calibration runs on the same scale.

The characteristic curve transformation method used the same calibrations (LOGIST runs) that were used in the concurrent method. However, all estimates of item parameters within a chain were placed on a common scale using a sequential transformation process developed by Stocking and Lord (1982). This procedure, which uses the common items (in this case, test forms) between two separate calibration runs, is based on the principle that, if estimates were error free, the proper choice of linear parameters (for placing item parameter estimates on the same scale) would cause the true scores on the common items from both calibrations to coincide. For example, the first two LOGIST runs for the ATP Biology chain produced parameter estimates for Forms BAC and UAC2 on one scale, and for Forms UAC2 and XAC on a different scale. Application of the characteristic curve transformation procedure yields a linear transformation, obtained from minimizing the difference between the true scores on items in Form UAC2 from the two calibrations, that is then used to place parameters for all items in the UAC2/XAC calibration on the same scale as those in the BAC/UAC2 calibration. The procedure is repeated, using the transformation obtained from the relationship between the true scores on Form XAC to place items from the XAC/TAC2 calibration on the same scale as the first two calibrations. The transformations were continued sequentially down each chain, resulting in a common scale for all item parameter estimates within a chain. The dotted lines between the boxes in Figure 2 indicate the common tests in each equating chain that were used to place item parameter estimates from the separate calibration runs on the same scale.

Once item parameter estimates on a common scale for two forms of a test were obtained, the relationship between estimated true scores on the two test forms was established in the following manner. The expected value of an examinee's observed formula score is defined as his or her true formula score. For the true formula score, $\xi$, we have

$$\xi = \sum_{i=1}^{n} \left[ \frac{(k_i+1)}{k_i} P_i(\theta) - \frac{1}{k_i} \right] , \tag{3}$$

where n is the number of items in the test form and $(k_i+1)$ is the number of choices for item i. If we have two test forms measuring the same ability $\theta$, then true formula scores $\xi$ and $\eta$ from the two tests are related by the equations

$$\begin{aligned} \xi &= \sum_{i=1}^{n} \left[ \frac{(k_i+1)}{k_i} P_i(\theta) - \frac{1}{k_i} \right] \\ \eta &= \sum_{j=1}^{m} \left[ \frac{(k_j+1)}{k_j} P_j(\theta) - \frac{1}{k_j} \right] \end{aligned} \tag{4}$$

Clearly, for a particular $\theta$ corresponding true scores $\xi$ and $\eta$ have identical meaning. They are said to be equated.

In practice, true formula score equating is carried out by substituting estimated parameters into equations (4). Paired values of $\xi$ and $\eta$ are then computed for a series of arbitrary values of $\theta$. Since we cannot know an examinee's true formula score, we act as if relationship (4) applies to an examinee's observed formula score.

For the concurrent equating method, item parameter estimates were only on the same scale for the two test forms that were calibrated together in the same LOGIST run (recall, each LOGIST run is represented in Figure 2 by a separate box). Therefore, the equating procedure (establishing the relationship between estimated true scores for two test forms) was applied sequentially, starting with the items calibrated in the first LOGIST run for each chain. Raw to scale conversion parameters were already available to convert raw scores on each of the initial test forms in the respective chains to the appropriate scale (i.e. College Board 200 to 800 or Graduate Record Examinations 200 to 990 scale). As an example of the sequential equating process, consider the ATP Biology test chain. Equivalent true formula score estimates were found for ATP Biology forms BAC and UAC2, resulting in a table of transformations of raw scores on UAC2 to the College Board scale. Form XAC was then equated to UAC2 resulting in a table to transformations for raw scores on XAC to the College Board scale. This procedure was repeated sequentially down the ATP Biology chain. The end product is a table of transformations of the raw scores on Form BAC to the College Board scale.

For the characteristic curve transformation method, a sequential equating procedure is not necessary because all item parameter estimates for the entire chain have been placed on the same scale. Only the equating of estimated true formula scores on the first form in the chain (parameter estimates obtained from the initial LOGIST run) to itself (parameter estimates obtained from the final LOGIST run) need be performed.

Assessment of Scale Drift

The amounts of scale drift attributable to the conventional and IRT equating methods were compared both graphically and analytically. Two types of graphical comparisons were made. First, graphs of final and initial (criterion) scaled score conversions were plotted for each equating method applied to each equating chain. Secondly, scaled score differences (final minus criterion) corresponding to raw scores on each of the four tests were plotted for each equating method. It should be noted that the equipercentile conversions do not extend over the entire raw score range for any of the tests. This is because it is only possible to obtain equipercentile conversions for scores that are actually observed in the equating samples. In practice, equipercentile conversion curves usually must be extrapolated in order to obtain scaled scores for all possible raw scores. Because extrapolation could possibly introduce an unknown source of error, no attempt to extrapolate the equipercentile equating results was made for this study.

In addition to the graphical comparisons, a discrepancy index was computed for each comparison of final and criterion scaled scores. For

example, for each raw score, x on GRE Biology Form SGR, there is a corresponding initial scaled score t and an estimated scaled score t' derived from one of the equating methods that was investigated. The smaller the difference, d, between t and t', the smaller the amount of scale drift and the more stable the equating method. A weighted mean square difference was used to summarize the difference between t and t'. The weighted mean square difference or total error can be broken down into the variance of the difference plus the squared bias, i.e.

$$\sum_j f_j d_j^2/n = \sum_j f_j (d_j-\bar{d})^2/n + \bar{d}^2 , \text{ or} \qquad (5)$$

(Total Error) = (Variance of Difference) + (Squared Bias)

where $d_j = (t'_j - t_j)$, $t'_j$ is the estimated scaled score for raw score $x_j$, $t_j$ is the initial or criterion scaled score for $x_j$, $f_j$ is the frequency of $x_j$, $n = \sum_j f_j$, $\bar{d} = \sum_j f_j d_j/n$, and the summation is over that range of x for which extrapolation of the equipercentile equating is unnecessary.

Summary statistics and discrepancy indices for each equating method applied to each equating chain were computed. The score frequencies used to compute the summary statistics and discrepancy indices were those for the total group taking the initial form of the test in each chain when the test was first administered.

Finally, in order to judge the importance of the results for the linear models, standard errors of the raw score-to-raw score equatings were computed for each of the equating chains. The computations were carried

out using the computer program AUTEST (Lord, 1975). Standard errors for the curvilinear equating methods (equipercentile and IRT) were not obtained because no method presently exists for determining the standard error of a chain of non-linear equatings. The standard errors were used to plot confidence intervals of plus and minus two standard errors around the final conversion lines for each of the linear equating methods applied to the respective equating chains.

Assessment of Goodness of Fit

Researchers often attempt to assess the fit of an item response theory model to real data using a chi-square test or other similar approaches (Wright and Panchapakesan, 1969; Wright and Stone, 1979). The problems associated with this approach have been discussed extensively in the literature (Gustafsson, 1980; Divgi, 1981; Rentz and Rentz, 1978; McKinley and Reckase, 1980). These problems have both theoretical and practical implications. From a theoretical point of view a problem exists in that chi-square tests require expected values that are available only when the parameters of the model ($\theta_k$, $a_i$, $b_i$ and $c_i$, in the case of the three-parameter model) are known; in actuality, we have only estimates of these parameters. These estimates are likely to behave differently from the known or true parameters in a statistical test. The practical problems are related to the interpretation of the chi-square values and their associated probability levels. One alternative to the various chi-square tests is the use of a graphical technique which involves the comparison of the regression of the observed proportion of people getting an item correct

on estimated $\theta$ (empirical regression) with the item response function based on the estimated item parameters (estimated regression) (Hambleton, 1980). The resulting plots are referred to as item ability regressions.

The problem with using item ability regression plots to assess goodness of fit is that the process is fairly subjective. The authors found it quite difficult to examine thousands of graphs (one for each item calibrated for the study) and make consistent judgements regarding the goodness of fit of each item. For this reason, it was decided to use a fit statistic leading to a chi-square like test in conjunction with the item ability regression plots. It should be emphasized that the statistic was used only to aid in the interpretation of the plots. No specific meaning was attached to either the size or the probability levels of the values obtained from the application of the statistic. The fit statistic and the item ability regression plots will each be described briefly in the remainder of this section.

The Fit Statistic

The fit statistic, referred to as $Q_1'$, is based on a statistic, $Q_1$, suggested by Yen (1981). The two statistics are very similar, the basic difference being the manner in which examinees are grouped into cells based upon their ability estimates. For both statistics, the initial step is to rank order examinees abilities. For $Q_1$, examinees are divided into 10 cells with approximately equal numbers of examinees in each cell. For $Q_1'$, examinees are divided into 17 cells, as follows. Examinees are placed into 15 equally spaced intervals for $\theta$ between +3 and -3. Those examinees with

$\theta$ greater than +3 are placed into a single cell and examinees with $\theta$ less than -3 are placed in another cell. Should any cell contain fewer than 5 examinees, it is collapsed with the adjacent cell closest to $\theta = 0$. The only remaining difference between the two statistics is that for $Q_1'$, the observed proportion of examinees in a particular cell is adjusted for examinees omitting the item. Using Yen's notation, the value of the fit statistic for item i is

$$Q'_{1i} = \sum_{j=1}^{17} \frac{N_j (O_{ij} - E_{ij})^2}{E_{ij}\,(1-E_{ij})} \quad , \tag{6}$$

where,

$N_j$ is the number of examinees in cell j, $O_{ij}$ is the observed proportion of examinees in cell j that passes item i (adjusted for omits) and, $E_{ij}$ is the predicted proportion of examinees in cell j that passes item i,

$$E_{ij} = \frac{1}{N_j} \sum_{k \varepsilon j}^{N_j} \hat{P}_i(\hat{\theta}_k) \tag{7}$$

where $\hat{P}_i(\hat{\theta}_k)$ is the item response function (equation 2) for item i. It should be noted that the summation is over examinees in cell j. The

degrees of freedom are the number of independent data points (cells) less the number of item parameters estimated from these data points. The number of estimated item parameters is not three in all cases. In some instances the value of the item discrimination parameter ($a_i$) was set to the upper bound for the a values.[2] In other instances the value of the psuedo-guessing parameter ($c_i$) was set to a common value.[3] Fit statistics were determined using $Q_1'$ for each of the achievement test items used in this study.

Item Ability Regression Plots

The item ability regression plots were obtained as follows. The ability scale ($\theta$) is subdivided into 15 equally spaced intervals for a range of -3 to +3. For each interval, equation (8) is used to compute $P_{ij}$, the proportion of people in interval j responding correctly to item i (adjusted for omits). That is,

$$P_{ij} = \frac{N_{ij}^{+} + N_{ij}^{o}/k}{N_{ij}} \quad , \text{ where} \qquad (8)$$

$N_j^+$ is the number of examinees in the jth interval responding correctly to item i,

$N_{ij}^o$ is the number of examinees in the jth interval that omitted item i,

k is the number of alternatives per item,

$N_{ij}$ is the number of examinees in interval j that reached item i.

---

[2]Upper and lower bounds were set for all item discrimination parameters to prevent the estimates from becoming unreasonably large or small.

[3]When LOGIST determines it cannot accurately estimate the c parameter for a certain item, due to insufficient information at lower ability levels, it uses an estimate of c obtained by combining all such items.

For each item, 15 P's are plotted as squares whose areas are proportional to $N_{ij}$ (these values constitute the empirical item ability regression). Also plotted with each square is a line of length $4\sqrt{PQ/N_{ij}}$, where P and Q are computed from the estimated item response function. The resulting 15 lines are centered on the estimated item response function which also appears on the plot. It should be noted that although the line is a rough estimate of the .95 confidence interval around the item response function, it is not being used as a statistical test for several reasons: (1) the use of the inappropriate symmetric normal approximation to the binomial confidence interval around the response function (particularly a problem for extreme values of P); (2) the use of an interval based on estimated item parameters; and (3) the use of 2 as a coefficient instead of 1.96. Item ability regression plots were obtained for each of the achievement test items used in this study.

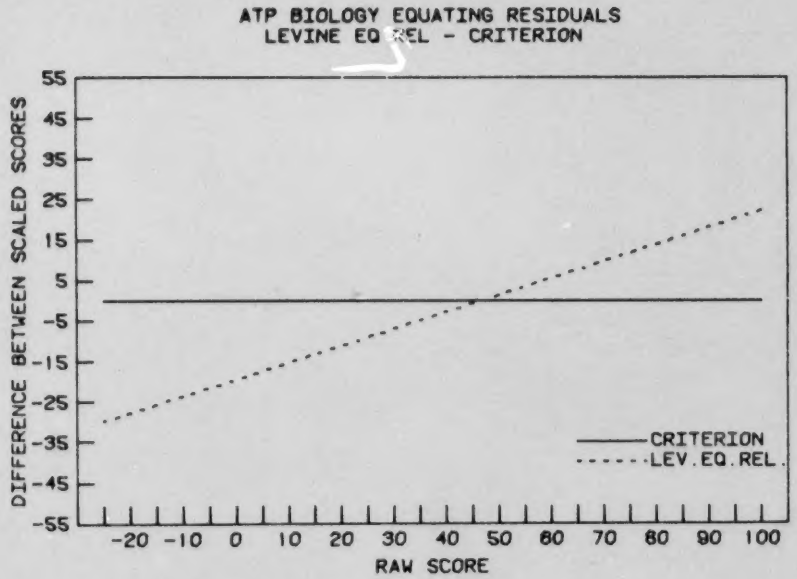
## Results

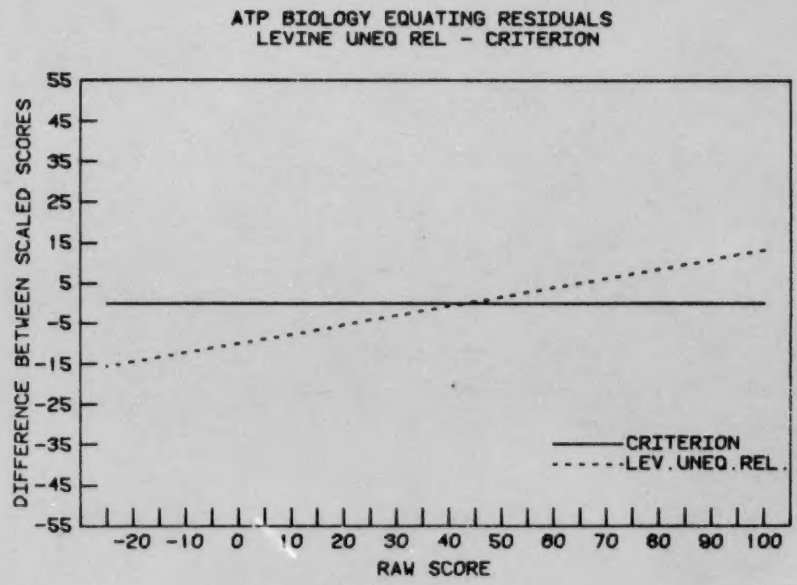
The initial (criterion) and final raw score to scaled score transformations for the first form in each achievement test equating chain (i.e., ATP Biology Form 3BAC, American History and Social Studies Form 3AAC, Mathematics Level II Form 3CAC2 and GRE Biology Form SGR) should be identical for all equating procedures. Departures resulting in scale drift may be due to sampling error and/or model fit problems.

The initial and final transformations resulting from the application of each equating method to each achievement test chain are given in Tables 1-4 of the Appendix. Also given in Tables 1-4 are the raw score frequencies

for the total group who took the first form of the test in each equating chain when it was introduced as a new test form. The information contained in Tables 1-4 is also presented graphically in Figures 1-4 of the Appendix. Although conversion tables and their accompanying plots (such as those presented in Tables 1-4 and Figures 1-4 of the Appendix) are informative, they tend to emphasize the similarities between the equatings rather than the differences. Tables and plots comparing equating residuals (such as Tables 5-8 of the Appendix and Figures 3-6 of the paper) allows finer distinctions to be made among the various equating methods applied to the respective achievement test chains.

Examination of Table 5 of the Appendix and Figure 3 of the paper, which summarize the equating residuals for the ATP Biology Test chain, indicates that both Levine linear methods had a tendency to overestimate the initial (criterion) scale values for the upper end of the score scale and to underestimate initial scale values for the lower end. Although the trends for the two Levine methods were similar, the Levine Equally Reliable method tended to produce greater discrepancies between the criterion and final conversions than the Levine Unequally Reliable Method. The Tucker linear method tended to overestimate the criterion scores for the entire range of raw scores. However, the discrepancies were generally less than those produced by either of the Levine methods. Of the three curvilinear methods (the two IRT methods and the equipercentile method), the equipercentile method produced the most discrepant scores. As expected, the greatest discrepancies for the equipercentile method occurred at the extremes of the





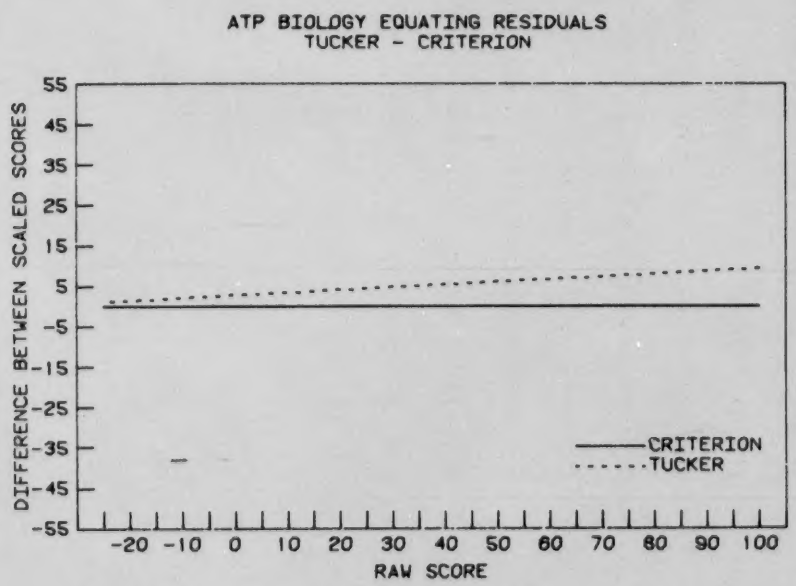
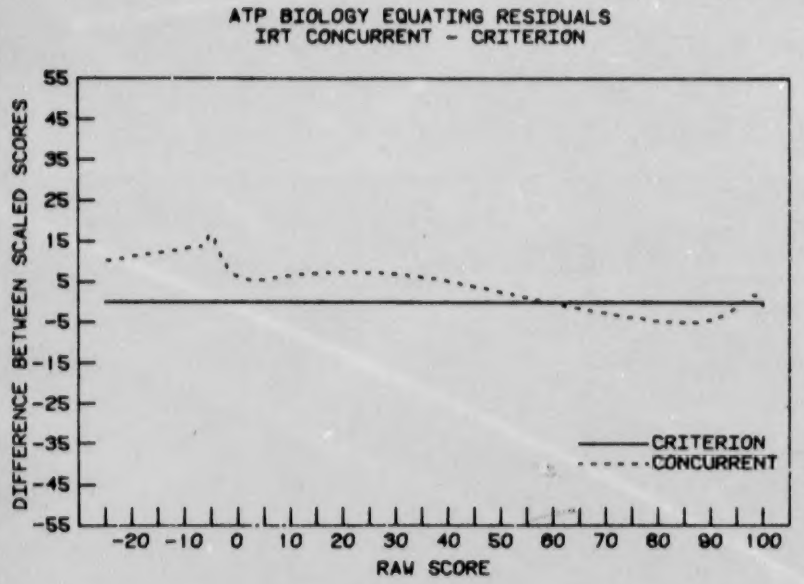
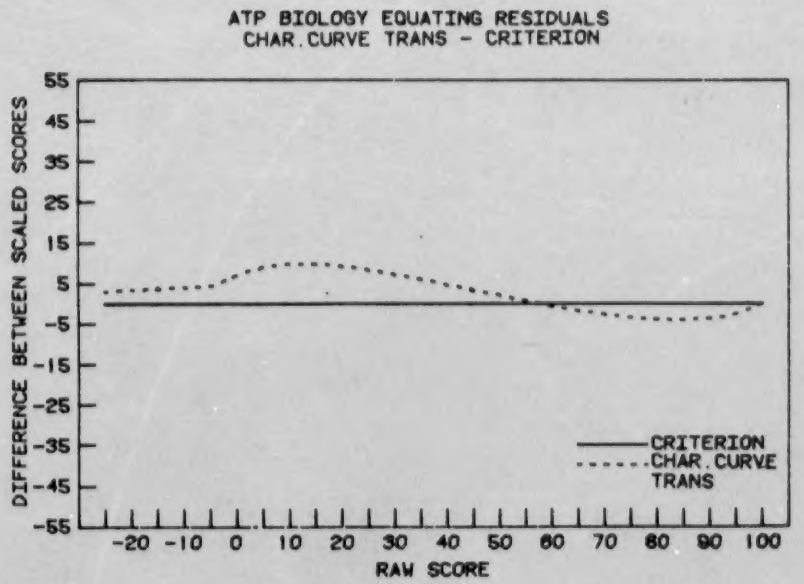








Figure 3: ATP Biology Equating Residuals.





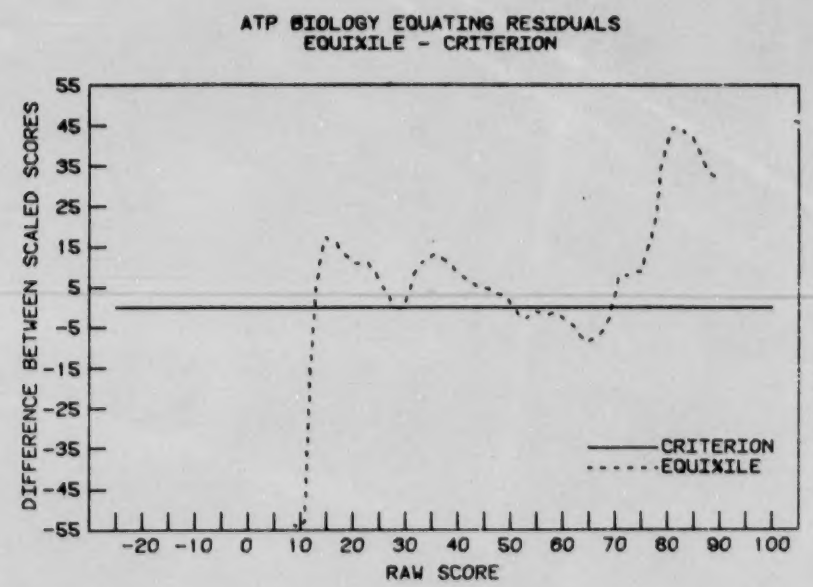





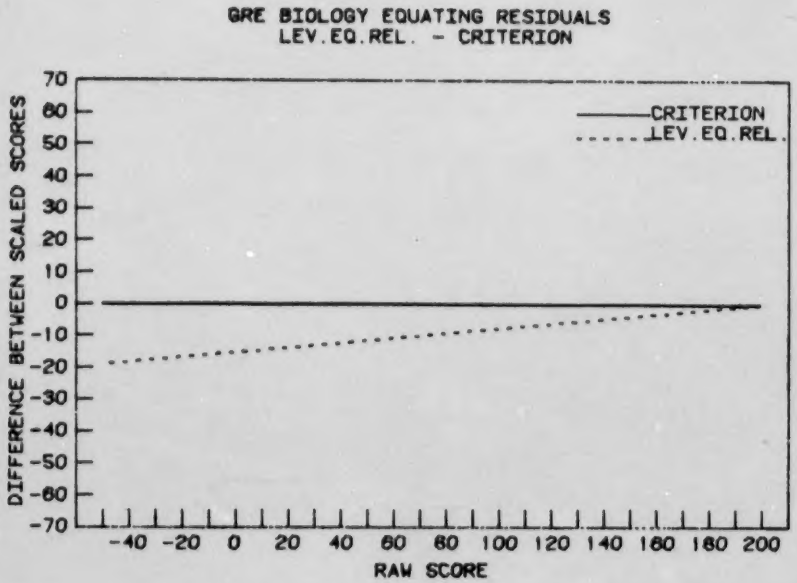


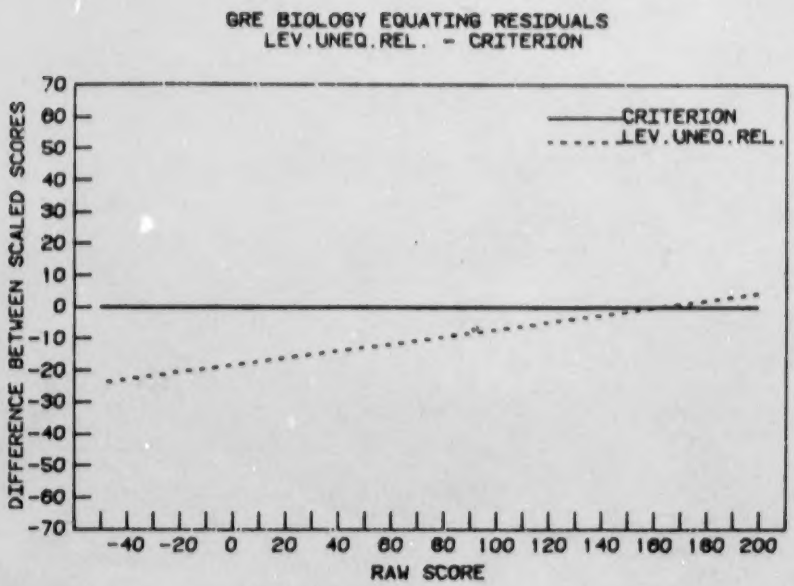

Figure 3 (cont.)

score scale. In general, the equipercentile method had a tendency to overestimate the criterion scores for most of the score reporting range. The results for the two IRT methods were very similar; both had a tendency to overestimate criterion scores at the lower to middle range of the score scale and to underestimate scores at the upper end.

Equating residuals for the GRE Biology test are presented in Table 6 of the Appendix and Figure 4 of the paper. Examination of the residuals indicates that all of the linear methods had a tendency to underestimate the criterion scores. Both of the Levine methods tended to underestimate scores in the lower portion of the score scale more than those in the upper portion. Exactly the opposite affect is observed for the Tucker method. There is a slight tendency for the Levine Unequally Reliable method to overestimate scores in the very upper end of the score distribution. The three curvilinear methods also had a general tendency to underestimate the criterion scores with the exception that the two IRT methods produced very slight overestimates of the criterion scores for a small range of scores in the upper end of the score scale.

A summary of the ATP Mathematics Level II equating residuals is presented in Table 7 of the Appendix and Figure 5 of the paper. It can be seen, from examination of this information, that the linear methods all underestimated the criterion scores in the upper end of the score scale and overestimated those in the lower end of the score scale. Of the three linear methods, the Tucker method resulted in the greatest discrepancies for scores in the upper and lower ends of the score scale. Similar to the




CR



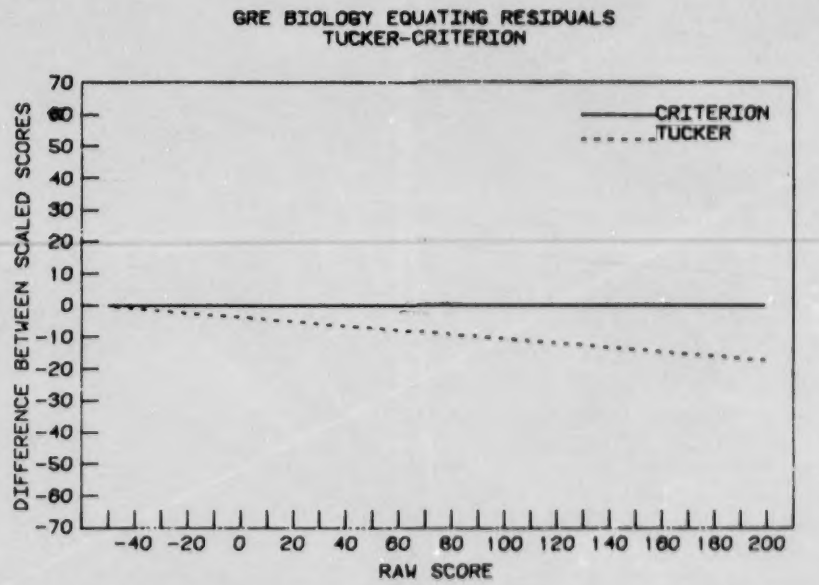
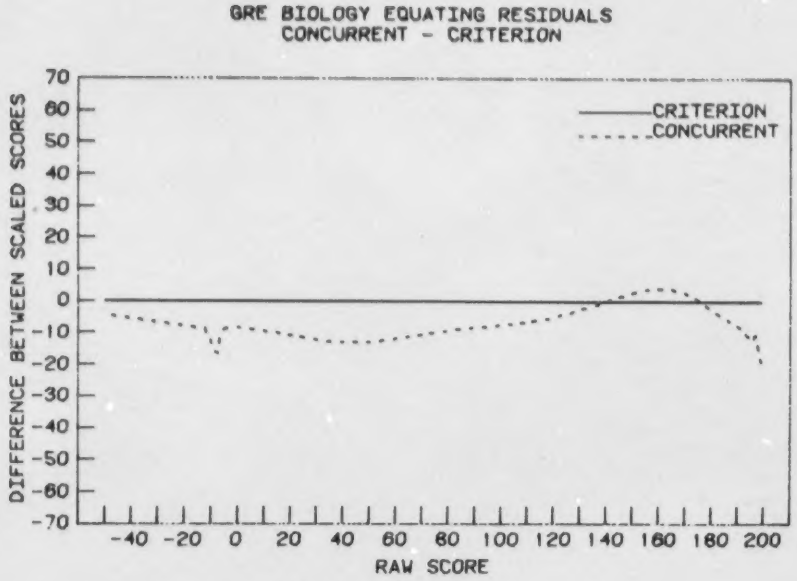
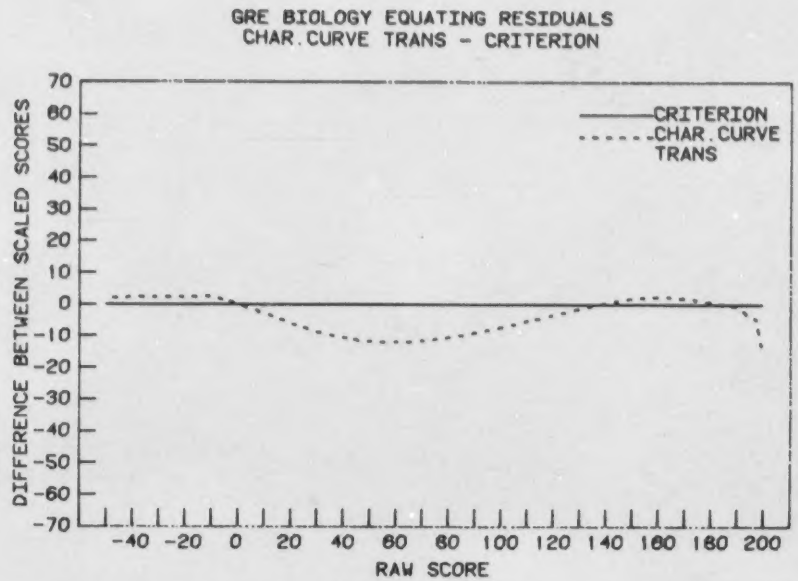







Figure 4: GRE Biology Equating Residuals.



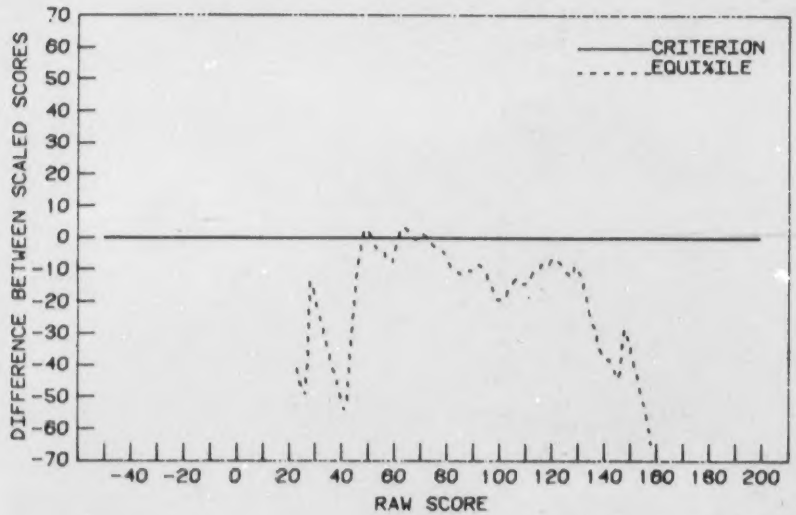
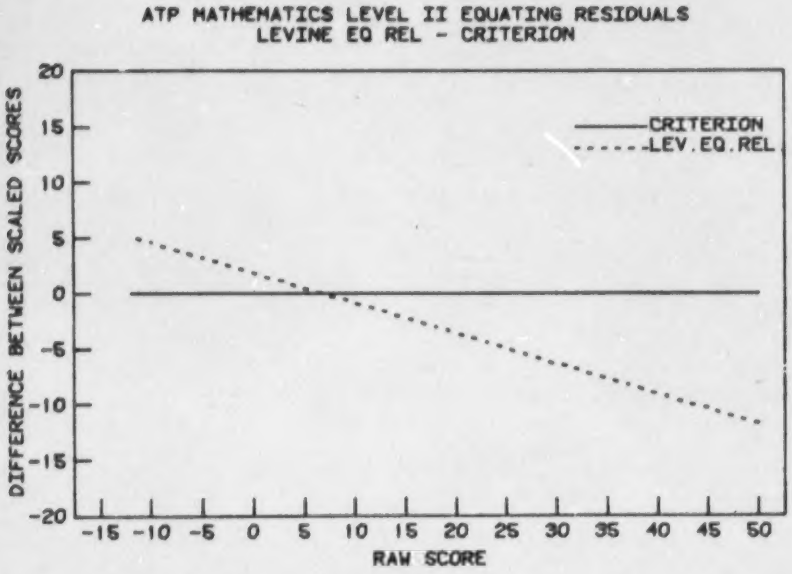



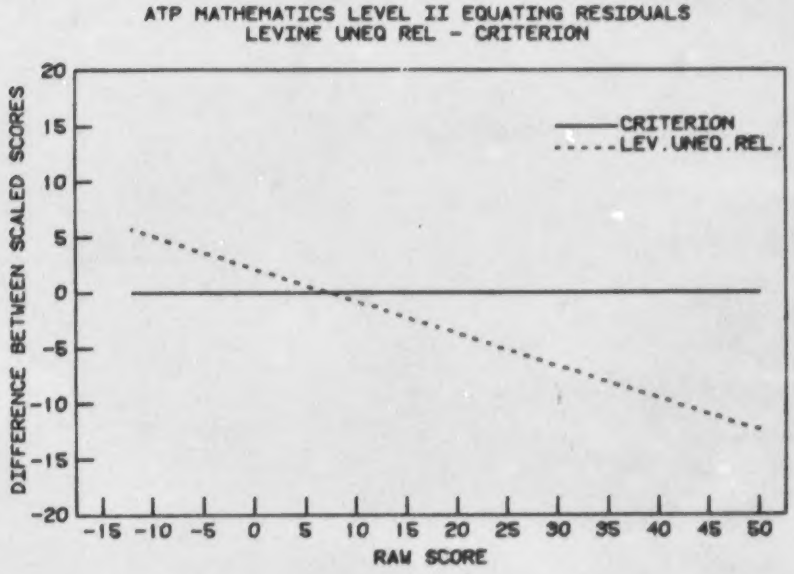







Figure 4 (cont.)



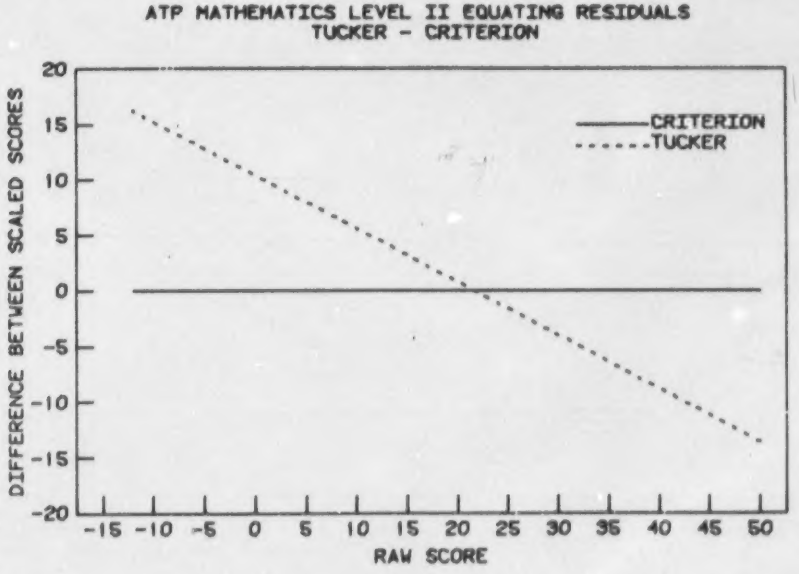
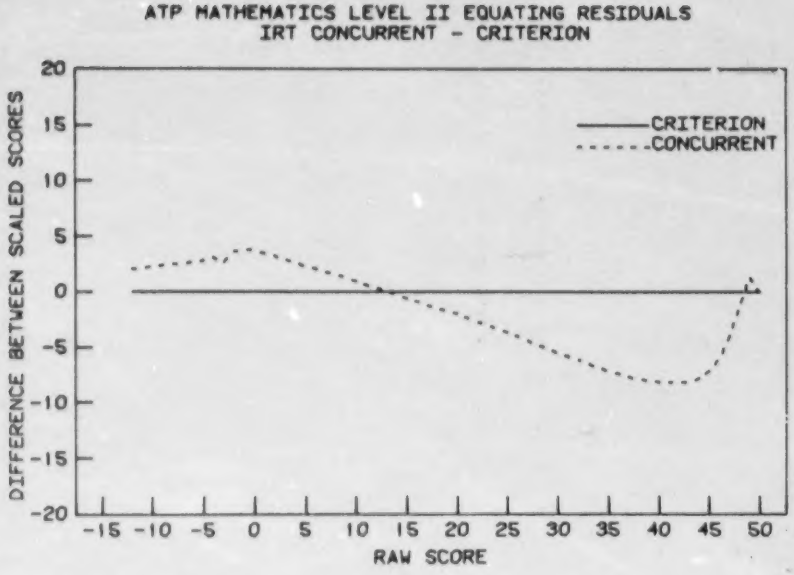



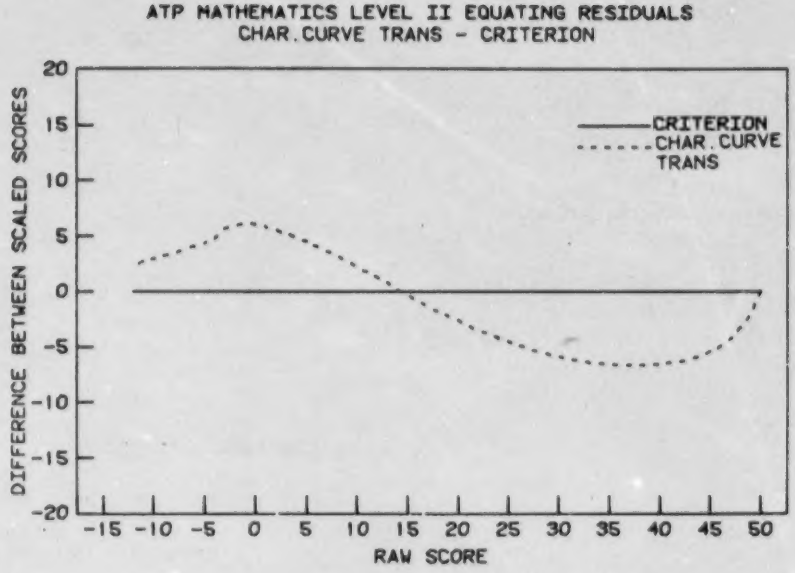







Figure 5: ATP Mathematics Level II Equating Residuals.



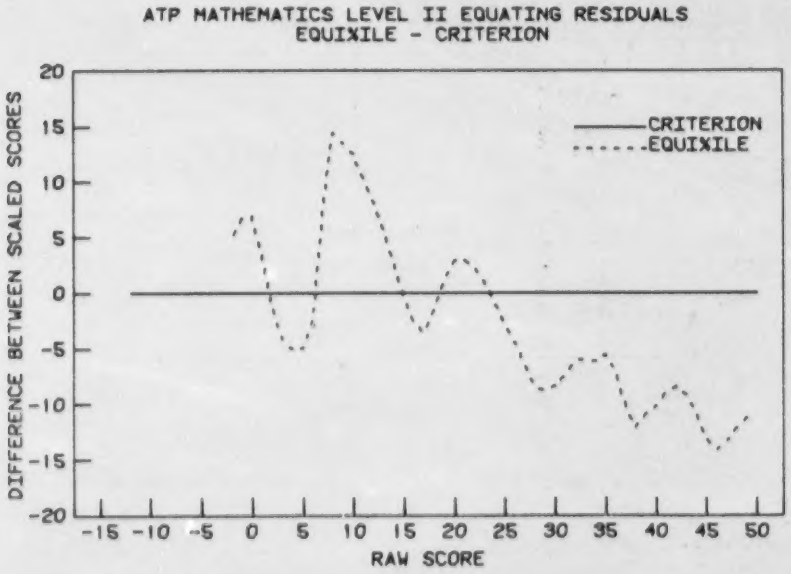







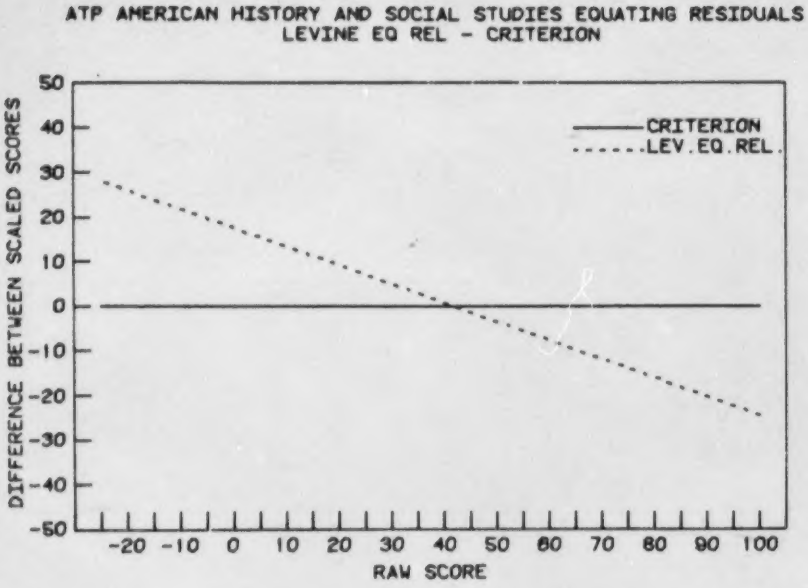


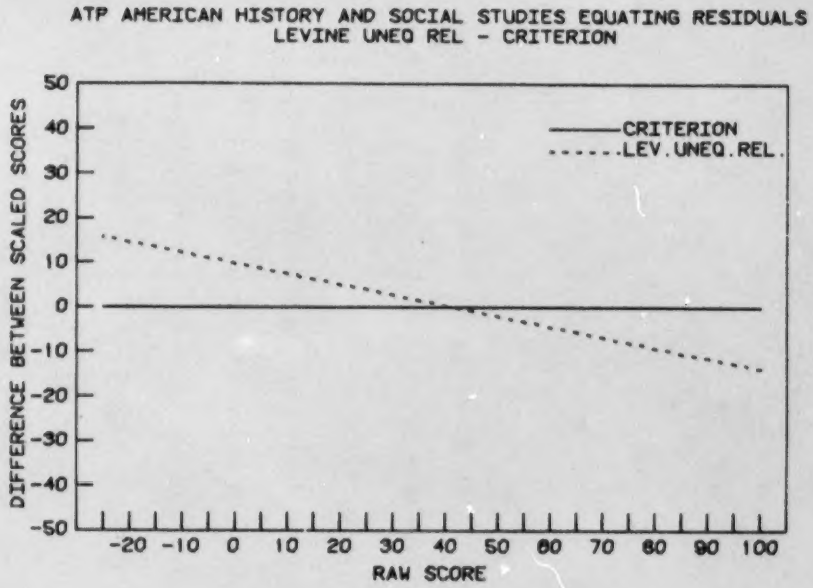

Figure 5 (cont.)

linear methods, the three curvilinear methods tended to overestimate lower criterion scores and underestimate criterion scores in the upper end of the score scale.

The American History and Social Studies Test equating residuals are presented in Table 8 of the Appendix and Figure 6 of the paper. Examination of the residuals for the linear equating methods indicates that all the methods overestimated lower criterion scores and underestimated higher criterion scores. It appears that, of the linear methods, the Levine Equally Reliable method produced the most discrepant scores for the extremes of the score scale. The IRT concurrent method showed a tendency to overestimate lower criterion scores and underestimate criterion scores in the upper end of the score scale. The IRT characteristic curve transformation method overestimated criterion scores in the low end of the score scale but showed remarkable agreement with criterion scores in the middle to upper end of the score scale. The equipercentile equating method had a tendency to overestimate criterion scores in the upper and lower ends of the score scale and underestimate criterion scores corresponding to raw scores that ranged from approximately 40 to 70.

The preceeding observations can be expanded upon through the examination of the summary statistics and discrepancy indices contained in Table 2 of the paper. The indices presented in this table have been described previously in the methodology section. Examination of the data for the ATP Biology Test indicates that the largest total error resulted from application of the equipercentile equating method and the smallest



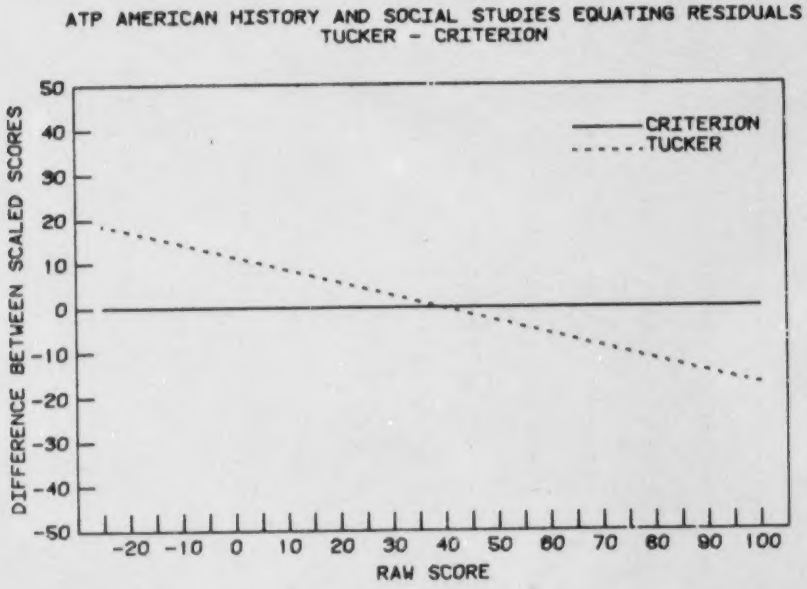
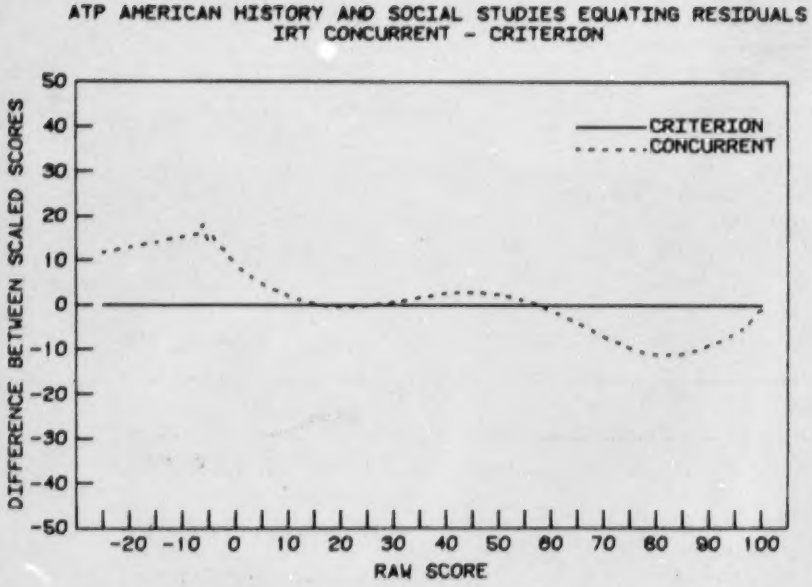
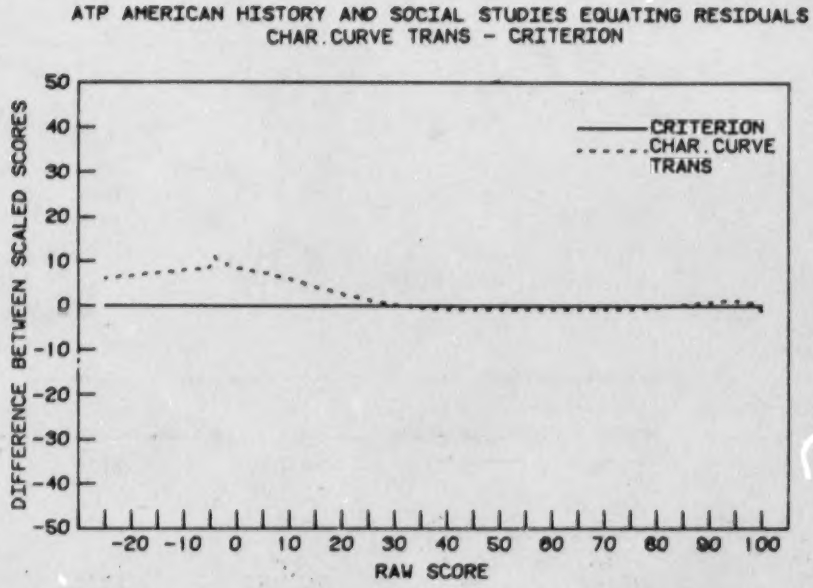








Figure 6: ATP American History and Social Studies Equating Residuals.



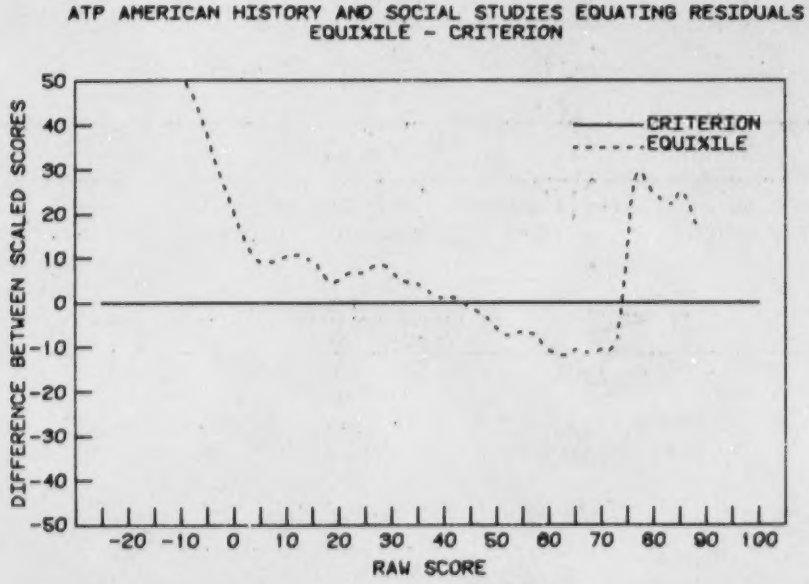










Figure 6 (cont.)

Table 2

Summary and Discrepancy Indices for Equating Methods
Used in the Achievement Test Scale Drift Study

| Index[a] | Initial Scale (Criterion) | Linear Equating | | | Curvilinear Equating | | |
|---|---|---|---|---|---|---|---|
| | | | | | | IRT Equating | |
| | | Tucker | Levine Eq. Rel. | Levine Uneq. Rel. | Equi%ile | Con-Current | Char. Curve Transf. |
| **ATP Biology** | | | | | | | |
| Scaled Score: | | | | | | | |
| Mean | 511.17 | 517.12 | 512.34 | 512.62 | 516.63 | 513.51 | 513.54 |
| Standard Dev. | 100.63 | 101.82 | 108.20 | 104.79 | 104.47 | 96.85 | 96.56 |
| Total Error[b] | | 36.74 | 58.56 | 19.41 | 184.54 | 20.60 | 22.50 |
| Bias | | 5.94 | 1.17 | 1.45 | 5.46 | 2.33 | 2.37 |
| S.D. of Difference | | 1.19 | 7.56 | 4.16 | 12.44 | 3.89 | 4.11 |
| **GRE Biology** | | | | | | | |
| Scaled Score: | | | | | | | |
| Mean | 629.63 | 619.84 | 621.10 | 621.30 | 619.05 | 621.28 | 621.45 |
| Standard Dev. | 109.84 | 108.11 | 111.75 | 112.70 | 106.68 | 112.90 | 113.04 |
| Total Error[b] | | 98.88 | 76.42 | 77.50 | 211.60 | 79.69 | 79.11 |
| Bias | | -9.79 | -8.53 | -8.33 | -10.58 | -8.35 | -8.18 |
| S.D. of Difference | | 1.73 | 1.91 | 2.85 | 9.99 | 3.16 | 3.49 |

[a]Computed for ATP Biology raw scores 7 through 89 (N=9080) and for GRE Biology raw scores 23 through 158 (N=3192).

[b]Total Error = $(\text{SD of Difference})^2 + (\text{Bias})^2$.

Table 2 (cont.)

Summary and Discrepancy Indices for Equating Methods
Used in the Achievement Test Scale Drift Study

| Index[a] | Initial Scale (Criterion) | Linear Equating | | | Curvilinear Equating | | |
|---|---|---|---|---|---|---|---|
| | | | | | | IRT Equating | |
| | | Tucker | Levine Eq. Rel. | Levine Uneq. Rel. | Equi%ile | Con-Current | Char. Curve Transf. |
| **ATP Mathematics Level II** | | | | | | | |
| Scaled Score: | | | | | | | |
| Mean | 650.13 | 648.93 | 645.42 | 645.24 | 647.62 | 646.75 | 646.81 |
| Standard Dev. | 82.94 | 78.28 | 80.30 | 80.11 | 78.18 | 80.16 | 80.05 |
| Total Error[b] | | 23.15 | 29.16 | 31.94 | 43.60 | 19.92 | 20.93 |
| Bias | | -1.21 | -4.71 | -4.89 | -2.51 | -3.39 | -3.33 |
| S.D. of Difference | | 4.66 | 2.64 | 2.83 | 6.11 | 2.91 | 3.14 |
| **ATP American History** | | | | | | | |
| Scaled Score: | | | | | | | |
| Mean | 470.79 | 470.71 | 471.68 | 471.22 | 472.28 | 471.25 | 471.08 |
| Standard Dev. | 91.88 | 87.06 | 84.85 | 87.93 | 87.96 | 90.37 | 90.44 |
| Total Error[b] | | 23.26 | 50.26 | 15.76 | 59.46 | 8.89 | 3.63 |
| Bias | | -.08 | .89 | .43 | 1.49 | .46 | .29 |
| S.D. of Difference | | 4.82 | 7.03 | 3.95 | 7.57 | 2.95 | 1.88 |

[a]Computed for ATP Mathematics Level II raw scores -2 through 49 (N=14744) and ATP American History scores -9 through 88 (N=18963).

[b]Total Error = $(\text{SD of Difference})^2 + (\text{Bias})^2$.

from the Levine Unequally Reliable method. Of the two IRT methods, the concurrent method resulted in slightly less total error than the characteristic curve transformation method. All of the methods tended to overestimate the criterion mean. All of the conventional methods also overestimated the criterion standard deviation. In contrast, the two IRT methods both underestimated the criterion standard deviation. The Levine Equally Reliable method gave the best estimate of the criterion mean and the worst estimate of the criterion standard deviation. The best estimate of the criterion standard deviation and the worst estimate of the criterion mean was given by the Tucker method. Bias accounted for over 90 percent of the total error for the Tucker method. In contrast, it accounted for less than 30 percent of the total error for the remaining methods. The methods that produced the most generally acceptable equating results were the Levine Unequally Reliable and the two IRT methods.

Inspection of the data for the GRE Biology Test presented in Table 2 indicates that the equating method resulting in the largest total error was the equipercentile method and that resulting in the smallest total error was the Levine Equally Reliable method. All of the methods underestimated the criterion mean. The two Levine methods and the two IRT methods overestimated the criterion standard deviation, whereas the Tucker and equipercentile methods underestimated the criterion standard deviation. For all of the methods, with the exception of the equipercentile method, at least 85 percent of the total error can be attributed to bias. Bias contributed to approximately 50 percent of the total error for the

equipercentile method. The most acceptable equating results were provided by the Levine and IRT methods, which all behaved very similarly.

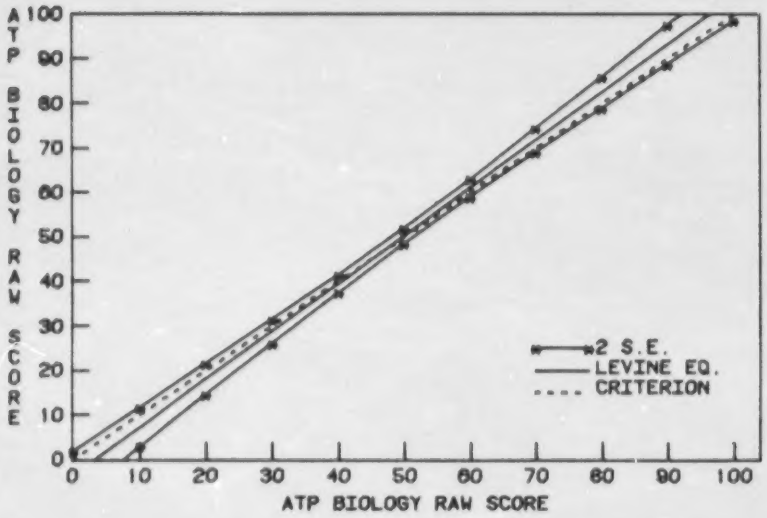
It can be seen, from the data presented in Table 2 for the ATP Mathematics Level II Test, that application of the IRT concurrent method resulted in the smallest total error whereas the equipercentile equating method resulted in the largest total error. All six equating methods underestimated the criterion mean and standard deviation. The worst estimate of the criterion mean was given by the Levine Unequally Reliable method and the best by the Tucker method. The Levine and IRT methods produced the best estimates of the criterion standard deviation. For both the equipercentile and Tucker methods, less than 20 percent of the total error can be attributed to bias. Bias contributed at least 75 percent to the total error for the two Levine methods and approximately 60 percent to the total error for the two IRT methods. The IRT equating results were very similar and, overall, the most acceptable.

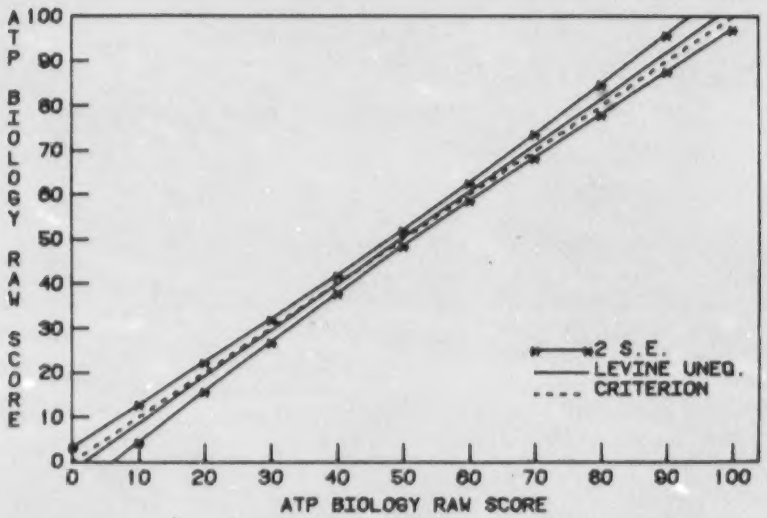
The data for the ATP American History and Social Studies Test presented in Table 2 shows that the smallest total error resulted from application of the IRT characteristic curve transformation method, and the largest total error from application of the equipercentile method. All methods underestimated the criterion standard deviation and, with the exception of the Tucker linear method, overestimated the criterion mean slightly. The Tucker method produced the best estimate of the criterion mean and the equipercentile method the worst; however, it should be noted that all methods produced very similar results. The two IRT methods gave the best

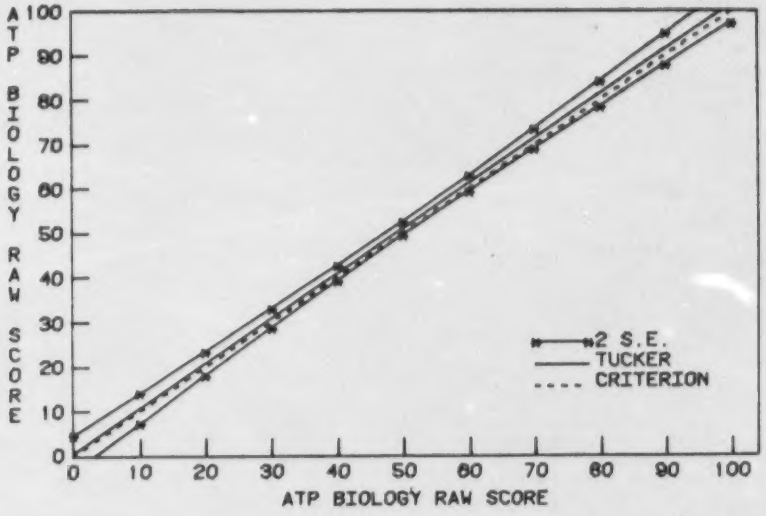
estimates of the criterion standard deviation and the Levine Equally Reliable method the worst. Bias accounted for less than 10 percent of the total error for all of the methods. Overall, the two IRT methods resulted in a remarkably small amount of scale drift.

In an effort to assess the importance of the discrepancies presented in Table 2, plots (in raw score units) were obtained of the final and criterion conversion lines for each linear equating method applied to the respective equating chains. For each plot, a confidence interval of plus and minus two standard errors (the method used to compute the standard errors is described in the methodology section) was drawn around the final conversion line. The plots are presented in Figures 7-10 of the paper. It is apparent, from examination of the plots, that no linear method applied to any equating chain resulted in converted scores that can be considered significantly different from the criterion scores.

Finally, although decisions regarding the feasibility of using IRT to equate the achievement tests investigated in this study should ultimately be based on assessments of scale drift, it was thought useful to attempt to evaluate the goodness of fit of the individual achievement test items to the three parameter logistic model. The method of assessment was basically judgemental and employed both the $Q_1$ statistic and the item ability regression plots described in the methodology section. The results of the goodness of fit assessment are presented in Table 3 of the paper. Examination of the data presented in Table 3 indicates that the average percentage of moderately poor to poorly fitting items ranges from a low of

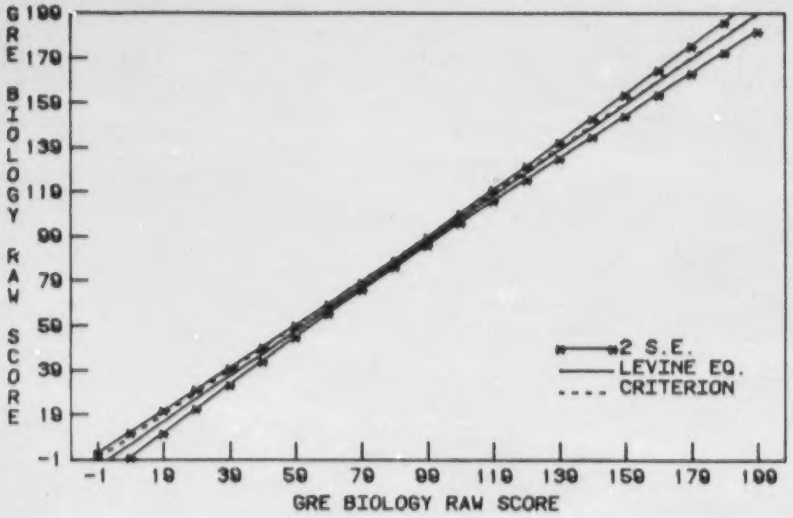


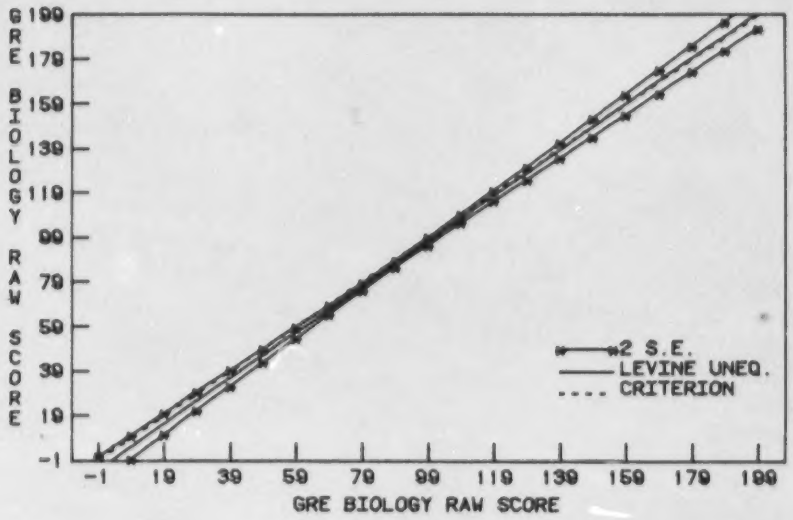






57 Figure 7: Plots of final and criterion conversion lines including confidence intervals of plus and minus two standard errors for all linear equating methods applied to the ATP Biology chain. 58



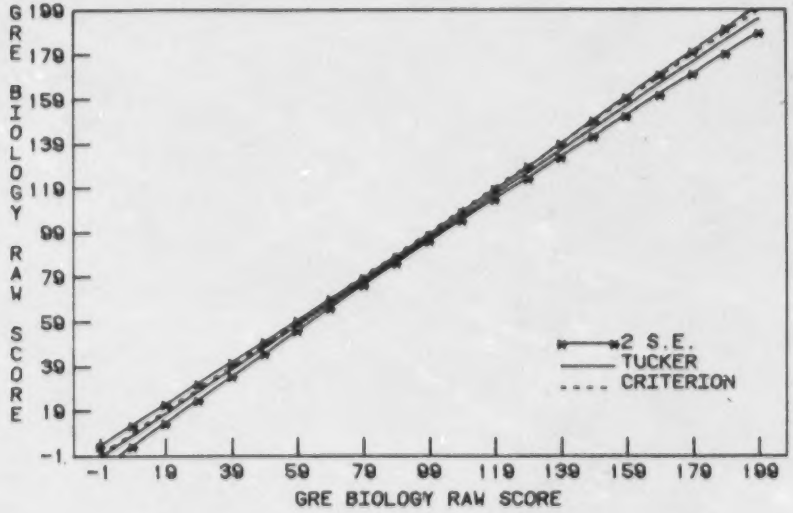
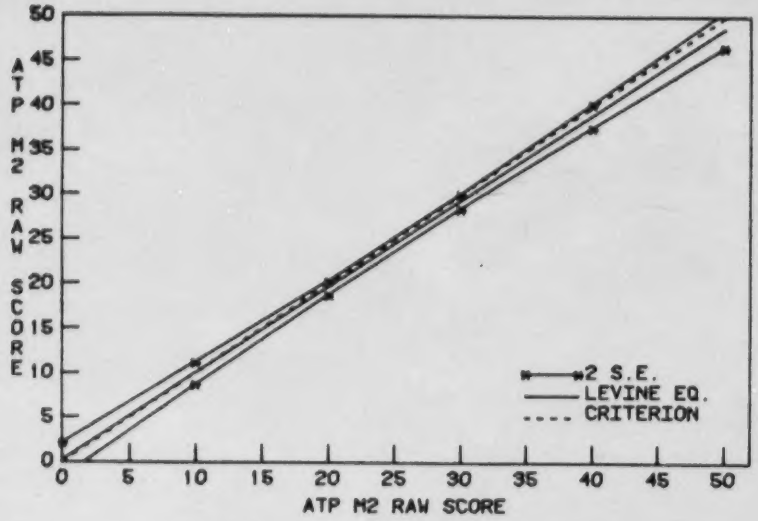
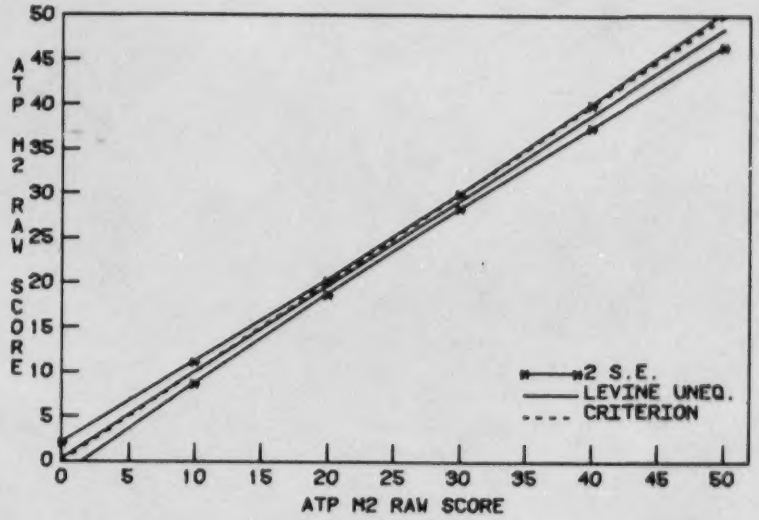








Figure 8: Plots of final and criterion conversion lines including confidence intervals of plus and minus two standard errors for all linear equating methods applied to the GRE Biology chain.



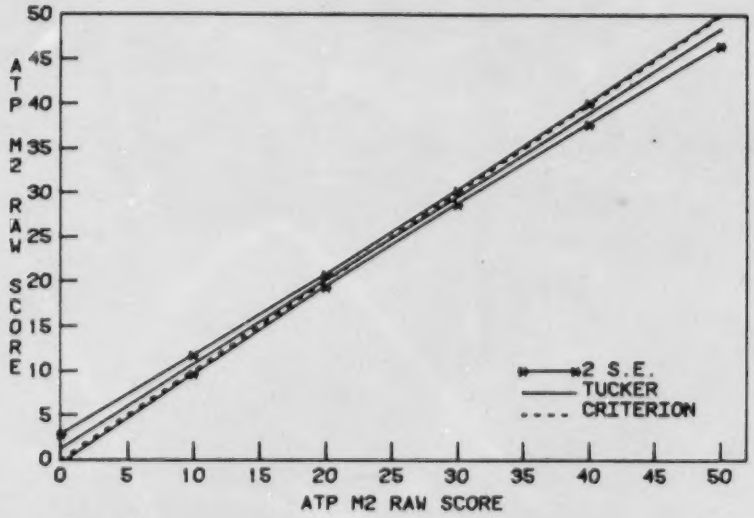
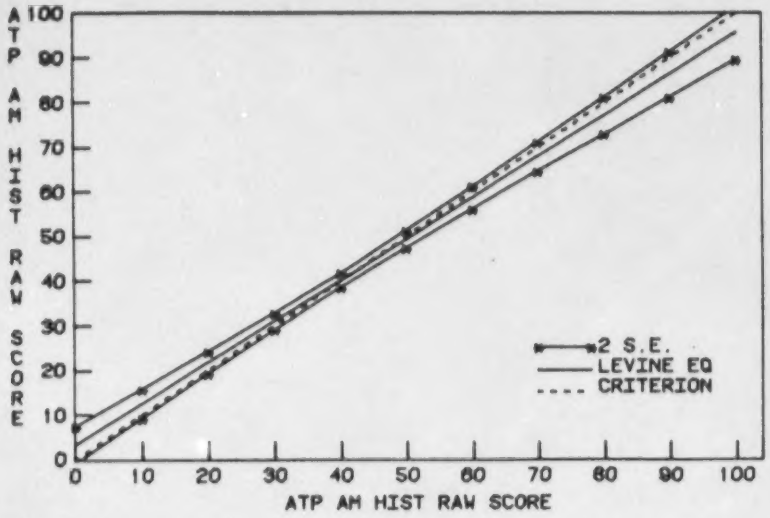
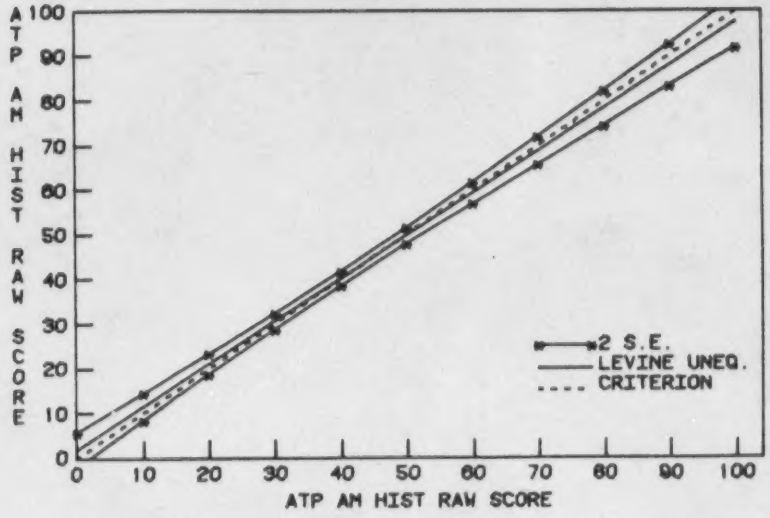




Figure 9: Plots of final and criterion conversion lines including confidence intervals of plus and minus two standard errors for all linear equating methods applied to the ATP Mathematics Level II chain.



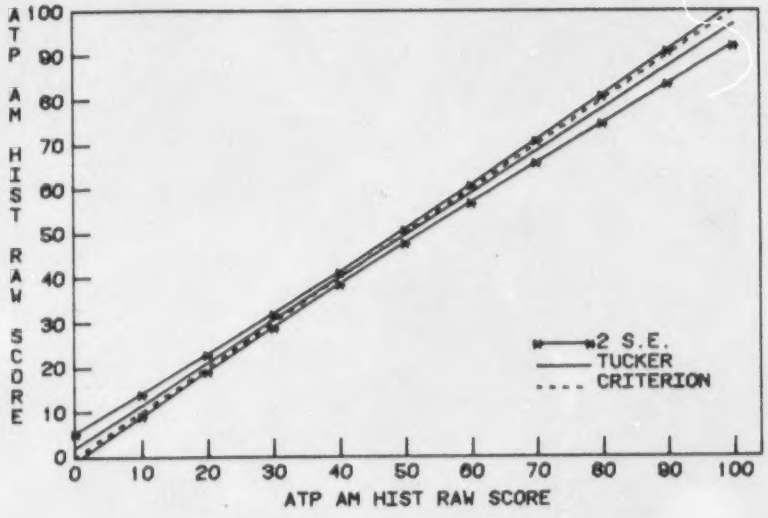






Figure 10: Plots of final and criterion conversion lines including confidence intervals of plus and minus two standard errors for all linear equating methods applied to the ATP American History and Social Studies chain.

Table 3

Numbers and Percentages of Items in Achievement Test Forms Judged as Having Moderately Poor to Poor Fit to the Three Parameter Logistic Model Using the $Q_1'$ Fit Statistic in Conjunction with Item Ability Regression Plots

| Form | Total Number of Items | Moderately Poor to Poorly Fitting Items[1] | | Poorly Fitting Items | |
|---|---|---|---|---|---|
| | | Number | Percentage | Number | Percentage |
| **ATP Biology Test** | | | | | |
| 3BAC | 100 | 9 | 9 | 3 | 3 |
| UAC2 | 100 | 18 | 18 | 8 | 8 |
| XAC | 100 | 15 | 15 | 6 | 6 |
| TAC2 | 100 | 22 | 22 | 19 | 19 |
| VAC1 | 100 | 8 | 8 | 5 | 5 |
| SAC2 | 100 | 19 | 19 | 14 | 14 |
| UAC1 | 100 | 10 | 10 | 5 | 5 |
| WAC | 100 | 11 | 11 | 8 | 8 |
| YAC | 100 | 9 | 9 | 3 | 3 |
| Total | 900 | 121 | 13.4[2] | 71 | 7.9[2] |
| **GRE Biology Test** | | | | | |
| SGR | 199 | 20 | 10 | 14 | 7 |
| K2-UGR1 | 210 | 19 | 9 | 8 | 4 |
| WGR | 210 | 20 | 10 | 4 | 2 |
| ZGR | 210 | 15 | 7 | 8 | 4 |
| XGR | 209 | 19 | 9 | 10 | 5 |
| K-UGR2 | 210 | 24 | 11 | 12 | 6 |
| Total | 1248 | 117 | 9.4[2] | 56 | 4.5[2] |
| **ATP Mathematics Level II Test** | | | | | |
| 3CAC2 | 50 | 6 | 12 | 4 | 8 |
| WAC | 50 | 6 | 12 | 3 | 6 |
| 3AAC | 50 | 7 | 14 | 5 | 10 |
| VAC1 | 50 | 8 | 16 | 5 | 10 |
| XAC | 50 | 4 | 8 | 2 | 4 |
| ZAC | 50 | 7 | 14 | 2 | 4 |
| 3BAC | 50 | 3 | 6 | 1 | 2 |
| Total | 350 | 41 | 11.7[2] | 22 | 6.3[2] |
| **ATP American History and Social Studies Test** | | | | | |
| 3AAC | 100 | 9 | 9 | 5 | 5 |
| XAC | 100 | 12 | 12 | 7 | 7 |
| UAC2 | 100 | 16 | 16 | 8 | 8 |
| YAC2 | 100 | 10 | 10 | 5 | 5 |
| K-WAC | 100 | 12 | 12 | 3 | 3 |
| YAC1 | 100 | 10 | 10 | 5 | 5 |
| Total | 600 | 69 | 11.5[2] | 33 | 5.5[2] |

[1]This category contains both those items judged to be poorly fitting and those judged to have moderately poor fit.

[2]The total number of moderately poor to poorly fitting items or the total number of poorly fitting items divided by the total number of items in the test forms evaluated.

9.4 for the GRE Biology equating chain to a high of 13.4 for the ATP Biology equating chain. The average percentage of poorly fitting items ranges from a low of 4.5 for the GRE Biology equating chain to a high of 7.9 for the ATP Biology chain. Given the narrow range of these average percentages, it would appear that no single equating chain can be singled out as having considerably better or poorer fitting items than any other chain.

## Discussion

### Conventional Equating Methods

As mentioned in the previous section, no linear equating method applied to any of the four achievement test equating chains produced scaled scores that can be considered seriously discrepant from the criterion scores. However, there are some differences among the results of the methods that are worth noting. The two Levine methods produced very similar estimates of the respective criterion means for the different equating chains; however, the estimates of the criterion standard deviations produced by these methods varied somewhat, particularly for the ATP Biology test and the ATP American History and Social Studies test. For both of these tests, the Levine Unequally Reliable model produced the better estimates of the criterion standard deviation. There is a fundamental difference between the two Levine models that has strong implications for their differential applicability to specific equating situations, i.e., the Levine Equally Reliable model is based on estimated means and standard deviations of observed scores whereas the Levine Unequally Reliable model is based on

estimated means and standard deviations of true scores. Lord (1980) states that in order to accurately equate two tests, i.e., produce scores on two tests such that it is a matter of indifference to examinees which test they take, the tests must be strictly parallel and perfectly reliable. Certainly, all of the tests used in this study depart somewhat from these criteria. It is difficult to predict how the various equating methods are effected by differences in test reliability; however, methods based on true score estimates, such as the Levine Unequally Reliable method (and also the IRT methods) should be least effected by this problem. It should be noted, however, that the two Levine methods performed very similarly when applied to the GRE Biology chain, the only chain containing test forms of different length and therefore, most likely, tests of differing reliability. One possible explanation is that the GRE Biology test forms are so long (199-210 items) that the differences in test length have only a negligible effect on the differences in test reliability. The Tucker method produced the best results of the three linear methods when applied to the ATP Mathematics Level II chain and the worst results of the three linear methods when applied to the GRE Biology chain. It produced better results than the Levine Equally Reliable method when applied to the ATP Biology chain and the ATP American History and Social Studies chain. The fact that the Tucker method performed reasonably well across all of the chains is worthy of further comment. Implicit to the derivation of the Tucker model is the assumption of random groups (Angoff, 1971, Levine, 1955). Since the samples for the test forms to be equated were not random samples from the

same administrations, and in some cases differed considerably in ability level (see Table 1 of the paper), it is quite surprising that the Tucker method gave such satisfactory results. Indeed, the fact that all of the linear methods produced conversions that could not be considered as significantly different from the criterion scores is quite surprising given that there is evidence of departures in parallelism between pairs of test forms that were equated in all of the equating chains (see Table 1). The lack of parallelism betwen test forms to be equated has particular implications for linear methods which require, in order to adequately describe the relationship between scores on two forms of a test, that the distribution of the scores differ only in their means and standard deviations. Lack of parallelism between two test forms generally results in a curvilinear relationship between raw scores, necessitating a curvilinear equating method to produce accurate results. It must be assumed, therefore, that the three linear equating methods investigated in this study are sufficiently robust both to departures in form to form parallelism and to differences in group ability of the degree exhibited by the four achievement test equating chains used for this study.

The equipercentile equating method produced the largest total error of all the equating methods applied to all the equating chains. A general problem with all equipercentile equating methods is that they are sensitive to scarcity of data in the extremes of the score distribution. The lack of stability of the equipercentile conversions provided for scores in the extremes of the score scale is quite apparent from inspection of the plots

given in Figures 3-6 of the paper. It should be noted that, for almost all the chains, the size of the total error for the equipercentile method can be attributed in large part to the standard deviation of the difference between the estimated and criterion scaled scores. In all cases, smoothing of the equipercentile conversions would have most likely produced a standard deviation of the difference more similar to that obtained for the linear models.

Item Response Theory Methods

The IRT concurrent and characteristic curve transformation methods gave very similar results when applied to the respective equating chains. On the one hand it could be concluded that this is not surprising, given that both methods used the same calibration (LOGIST) runs and that the number of common items used to link the separate calibration runs for the characteristic curve transformation method was quite large, ranging from 50 items for the ATP Mathematics Level II chain to 210 items for most of the forms in the AE Biology chain. On the other hand, the two methods employ fundamentally different processes (as described in the methodology section) to arrive at the final converted scores that were compared to the criterion scores. Considering the basic procedural differences between the two methods, it is quite surprising that they produced results which were in such close agreement when applied to all the equating chains. The results obtained for the IRT methods employed in this study can be compared to those obtained in a similar study conducted by Petersen, Cook, and Stocking (in press). Petersen, et al., used scale drift as the criterion to compare

the application of several equating methods, including the IRT concurrent and IRT characteristic curve transformation methods, to the verbal and mathematical sections of the Scholastic Aptitude Test (SAT). The two IRT methods did not perform similarly when applied to either the verbal or mathematical aptitude test data. In both cases, the IRT concurrent method produced more acceptable equating results. The number of linking items used for the characteristic curve transformation method applied to the SAT data was considerably less than the number employed for all of the equating chains used in the present study. Thus a plausible explanation for the close agreement between the two methods, as applied to the respective achievement test equating chains, might be the large number of common items used to link parameter estimates from the separate LOGIST runs.

The most notable observation that can be made regarding the two IRT methods employed in this study is that they both produced very acceptable equating results for all of the tests that were investigated. Either the IRT methods used in this study are robust to violations of the assumption of unidimensionality or the particular achievement tests studied are more unidimensional than a review of the multiple content areas they are purported to measure would lead one to believe. Most likely both of these factors are contributing to the equating results. Since it is highly unlikely that any test of aptitude or achievement is truly unidimensional and since IRT methods have been used successfully to equate a variety of different types of tests (see Cook and Eignor, 1983, for a comprehensive review of IRT equating studies), it seems reasonable to assume that IRT

equating methods are somewhat robust to violations of the assumption of unidimensionality.

On the other hand, one of the requirements underlying all of the equating models used in this study, if Lord's (1980) equity requirement is to be met (see page 36), is that the two tests to be equated are unidimensional (Morris, 1982). Because IRT equating models assume unidimensionality on the item level whereas the linear and equipercentile models used for this study only assume unidimensionality at the test score level, one might expect violations of this assumption to have a more serious effect on the IRT equating results. However, unidimensionality is a necessary condition for the establishment of a single common metric regardless of the equating model. Given the application of the linear equating methods did not produce converted scores that could be considered significantly different from the criterion scores, it is probably reasonable to assume that all of the achievement tests investigated in this study are approximately unidimensional, at least on the total score level.

The goodness of fit assessment was conducted in the hope that if application of the IRT methods to a particular achievement test equating chain produced seriously discrepant results, the results might be explained by lack of fit of the items for the particular test to the three-parameter logistic model. As mentioned previously, all of the tests contained a certain percentage of items which were judged to fit the model poorly. Apparently the equating process is robust, to a certain extent, to the lack of fit of individual items, at least to the extent of the lack of fit observed for these data.

Comparison of IRT and Conventional Methods

For all equating chains, the IRT methods produced less total error than either the Tucker or Equipercentile equating methods. For the ATP Mathematics Level II chain and the ATP American History and Social Studies chain, the IRT methods resulted in less total error that any of the other equating methods employed. For the ATP Biology chain, the Levine Unequally Reliable method produced a slightly smaller total error than either of the IRT methods. Finally, for the GRE Biology chain, both of the Levine methods resulted in slightly less total error than either of the IRT methods.

These comparisons can be viewed from several points of view. The fact that all methods, with the exception of the equipercentile method, provided fairly similar and reasonable equating results is comforting in that it provides evidence of the viability of the conventional linear methods that have been used historically to equate the tests. The comparisons also indicate that IRT methods provide a reasonable alternative to the conventional methods, should there be a particular need to use them. For example, if the specifications for one of the tests were revised sufficiently such that it was anticipated that the relationship between a new form of the test and the form it was equated to might be curvilinear, it appears as though either IRT method employed in this study would provide an effective method of estimating the curvilinear relationship.

## Conclusions

The results of this study indicate that it is feasible to use item response theory to equate the four achievement tests selected for investigation. The results also indicate that the conventional linear methods typically used to equate the tests perform quite adequately. The question of whether the IRT methods used in this study are sufficiently robust to violations of the assumption of unidimensionality, or whether achievement tests, of the type used in this study, give rise to sufficiently unidimensional data, must be resolved before the results of the study can be generalized to other achievement testing situations. Of fundamental importance is the development of a methodology that can be used to determine the number of underlying dimensions measured by a set of test items (see Cook, Dorans, Eignor and Petersen, 1983, for a description of an initial attempt at establishing a methodology). If the number of dimensions included by the various achievement tests used in this study could be ascertained, it would be possible to make a statement regarding the robustness of the two IRT methods and to generalize the results of the study to other achievement tests that exhibit similar dimensionality.

## References

Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Cook, L. L., and Eignor, D. R. Practical considerations regarding the use of item response theory to equate tests. In R. K. Hambleton (Ed.), Applications of item response theory. Vancouver, B.C.: Educational Research Institute of British Columbia, 1983.

Cook, L. L., Dorans, N., Eignor, D. R., and Petersen, N. S. An assessment of the relationship between the assumption of unidimensionality and the quality of IRT true-score equating. Paper presented at the annual meeting of AERA, Montreal, 1983.

Divgi, D. R. Does the Rasch model really work? Not if you look closely. Paper presented at the annual meeting of NCME, Los Angeles, 1981.

Gustafsson, J-E. Testing and obtaining fit of data to the Rasch model. British Journal of Mathematical and Statistical Psychology, 1980, 33, 205-233.

Hambleton, R. Latent ability scales: Interpretation and uses. In S. Mayo (Ed.), New Directions for Testing and Measurement; Interpreting Test Performance, No. 6. San Francisco: Jossey-Bass, 1980.

Levine, R. E. Equating the score scales of alternate forms administered to samples of different ability (RB-55-23). Princeton, N.J.: Educational Testing Service, 1955.

Lord, F. M. Estimation of latent ability and item parameters when there are omitted responses. Psychometrika, 1974, 39, 247-264.

Lord, F. M. Automated hypothesis tests and standard errors for nonstandard problems. The American Statistician, 1975, 29, 56-59.

Lord, F. M. Applications of item response theory to practical testing problems. Hillsdale, N.J.: Erlbaum, 1980.

McKinley, R. L., and Reckase, M. D. A comparison of the ANCILLES and LOGIST parameter estimation procedures for the three-parameter logistic model using goodness of fit as a criterion. Research Report 80-2. Arlington, VA: Personnel and Training Research Programs, ONR, 1980.

Morris, C. N. On the foundations of test equating. In P. W. Holland and D. B. Rubin (Eds.) Test equating. New York: Academic Press, 1982.

Petersen, N. S., Cook, L. L., and Stocking, M. L. IRT versus conventional equating methods: A comparative study of scale stability. Journal of Educational Statistics, in press.

Rentz, R. R., and Rentz, C. C. Does the Rasch model really work? A discussion for practitioners. ERIC Report No. 67. Princeton, NJ: Educational Testing Service, 1978.

Stocking, M. L., and Lord, F. M. Developing a common metric in item response theory. Research Report RR-82-5-ONR. Princeton, N.J.: Educational Testing Service, 1982.

Wingersky, M. S. LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (Ed.), Applications of item response theory. Vancouver, B.C.: Educational Research Institute of British Columbia, 1983.

Wingersky, M. S., Barton, M. A., and Lord, F. M. Logist V user's guide. Princeton, N.J.: Educational Testing Service, 1982.

Wright, B. D., and Panchapakesan, N. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.

Wright, B. D., and Stone, M. H. Best test design. Chicago, IL: Mesa Press, 1979.